UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE INGENIERÍA HIDRÁULICA Y AMBIENTAL

Programa Doctoral en Ingeniería del Agua y Medio Ambiental

INCIDENCIA DE LA CALIDAD EL AIRE EN EL DESARROLLO URBANO SOSTENIBLE. METODOLOGÍA DE PRONÓSTICO BASADO EN HERRAMIENTAS DE APRENDIZAJE AUTOMÁTICO

Tesis doctoral

Abril, 2021

Nidia Isabel Molina Gómez
Autora

P. Amparo López Jiménez
Directora

Página intencionalmente en blanco.

Agradecimientos

La familia es el soporte de la vida, está presente para dar tus primeros pasos y somos afortunados quienes nos sentimos acompañados en caminos tan importantes como lo es la elaboración de un trabajo doctoral. Quiero agradecer a mis padres y a mis hermanas por su acompañamiento incondicional; en todo mi camino han sido mi gran compañía y su espaldarazo ha sido fundamental en cada etapa de mi vida.

Mi directora P. Amparo López Jiménez, quien estuvo presente acompañando y orientando el proceso a través de su experiencia, y sus palabras que siempre fueron de impulso y continuidad; “sin prisa, pero sin pausa” de las mejores frases en nuestras conversaciones de dirección que retumbaron y se mantuvieron. Sus aportes, sus observaciones y su guía fueron clave en mi formación y desarrollo de este trabajo.

También quiero agradecer a José Luis Diaz Arévalo, quien desde su experiencia me orientó en elementos claves para avanzar con pasos firmes.

Quiero agradecer a mis compañeros Claudia Navarrete, Karina Solano y Darwin Mena, con quienes inicié esta etapa doctoral, con quienes nos acompañamos en el proceso y compartimos algunas experiencias, gracias por su apoyo e impulso para iniciar este camino y finalizarlo.

Agradezco especialmente a mis amigos, a Dayam Calderón por su acompañamiento en los aprendizajes, a Ronal Sierra por estar presente para apoyarme, y a mis amigos de la Universidad Santo Tomás la segunda familia, por sus palabras de impulso y de apoyo para generar aportes en nuevo conocimiento.

Agradezco a la Universitat Politècnica de València y a la Universidad Santo Tomás. Agradezco a cada compañero que crucé en el desarrollo de esta etapa porque alguna palabra, observación, sugerencia en cualquier charla me incentivó a continuar y con toda seguridad, contribuyó para cerrar cada eslabón en la generación de valor a la idea inicialmente propuesta.

Agradezco a cada persona que hizo parte del equipo en la recopilación de información, así como a las entidades que la facilitaron; agradezco además a los revisores de las publicaciones, quienes desde su experticia aportaron con sus requerimientos para mejorar cada producto. Me excuso si no he mencionado a alguna persona, pues a lo largo de este camino lleno de aprendizajes interactué con diferentes profesionales. A todos muchas gracias, por su apoyo y compañía.

Página intencionalmente en blanco

Resumen

La calidad del aire es un determinante de la salud y bienestar de las poblaciones que, de acuerdo con cifras de la Organización Mundial de la Salud, ha cobrado anualmente cerca de siete millones de vidas humanas. Es además un factor clave para el avance de las naciones y de los territorios en sus diferentes escalas, pues es el ser humano el eje principal de las decisiones en materia de desarrollo sostenible. Al respecto, en el marco de la política pública a nivel mundial, se han diseñado y puesto en marcha una serie de elementos clave, cuyo fin ha sido superar los desafíos que impone el crecimiento poblacional y sus consecuentes requerimientos; se trata de grandes hitos como los Objetivos de Desarrollo del Milenio y en la actualidad, los Objetivos de Desarrollo Sostenible (ODS) con la Agenda 2030. Estos elementos se han venido implementando con el soporte de acuerdos, convenios, protocolos, planes, programas y proyectos que desde el ámbito de país son direccionados y decantados para su implementación.

Existen reportes nacionales de avance sobre la implementación de metas específicas, según la agenda de cada país y en algunos casos en el ámbito de ciudad, cuyos indicadores pueden integrarse en las dimensiones centrales y más conocidas del desarrollo sostenible: la dimensión ambiental, la social y la económica. Se destaca que existe información sobre el monitoreo del estado de la calidad de los recursos y de las condiciones específicas del territorio en diversos temas. Sin embargo, no en todos los territorios, en sus diferentes escalas espaciales, se realiza continua evaluación de su desempeño sostenible y, además factores de deterioro ambiental como la contaminación del aire, son tratados como determinantes aislados con la generación de reportes de su comportamiento y el desarrollo de planes de monitoreo y de mitigación.

Para los diferentes temas que hacen parte de las dimensiones de la sostenibilidad, existen herramientas de modelación para evaluar el comportamiento de sus indicadores; sin embargo, no se cuenta con un instrumento que pronostique el nivel de avance en el desarrollo sostenible y además que identifique la influencia de la calidad del aire en su comportamiento. Las herramientas de aprendizaje automático pueden aportar en la respuesta a dicha situación, dado que consisten en una estructura útil para el pronóstico del comportamiento de un conjunto de datos, a partir de unos objetivos de regresión y/o clasificación y que soportan con sus resultados a la toma de decisiones.

Por consiguiente, el objetivo central de este trabajo doctoral es establecer la incidencia de la calidad del aire sobre el desarrollo urbano sostenible, en sus dimensiones ambiental, social y económica, mediante el uso de herramientas de aprendizaje automático, como soporte para la toma de decisiones. Este objetivo involucra el diseño y ejecución de una metodología para identificar la influencia de indicadores en materia de calidad del aire, sobre el desarrollo urbano sostenible. Para su logro se han definido cinco objetivos específicos, que consideran la relevancia de los micro territorios como escenario geográfico fundamental al logro de los ODS.

Este trabajo doctoral se lleva a cabo a partir de un estudio de caso en una localidad de la ciudad de Bogotá, en Colombia que es la capital del país, con una extensión de 350 km^2, situada sobre una planicie altitudinal en la cordillera oriental y a 2625 metros sobre el nivel del mar. Bogotá es una de las ciudades más pobladas en América Latina con 8.3 millones de habitantes y es una de las capitales mundiales que ha presentado altos niveles de contaminación por material particulado, siendo éste un factor de riesgo para su población.

Los siguientes son los objetivos específicos planteados para el desarrollo de esta investigación:

- Revisar el estado del arte para la identificación de las variables y parámetros que podrían calificar las dimensiones individuales del desempeño sostenible en función del componente atmosférico y calidad del aire, así como la escala temporal para el análisis
- Realizar un análisis estadístico de desempeño sostenible de acuerdo con las variables y parámetros seleccionados para una localidad específica en la ciudad de Bogotá
- Identificar y seleccionar las herramientas de aprendizaje automático aplicables
- Aplicar las herramientas de aprendizaje automático seleccionadas para determinar la incidencia de la calidad del aire sobre el desarrollo urbano sostenible en una localidad de la ciudad de Bogotá y,
- Obtener conclusiones acerca del análisis y proponer desarrollos futuros

Este trabajo inició con el conocimiento de localidad de Kennedy en términos de calidad del aire y salud, identificando el estado del arte en dichos factores; se analizó el comportamiento de contaminantes atmosféricos y variables meteorológicas en el periodo 2009 a 2017, el comportamiento de enfermedades del sistema respiratorio, según reportes entregados por entidades gubernamentales, y otros factores que incluyen la proximidad de los hogares a vías primarias y secundarias, además de los resultados de un trabajo de campo realizado en 2017. Con la aplicación de las herramientas de aprendizaje automático denominadas Random Forest y Adaptative boosting, así como información de entrada en formato ráster, estructurada a partir de una composición de bandas con el uso del software ArcGis y el lenguaje de programación de acceso libre R, se determinaron las zonas de la localidad con mayor probabilidad de presentar enfermedades del sistema respiratorio, a la vez que las principales variables que determinaron dicha condición en el territorio.

Posteriormente, a partir de una revisión y análisis de diferentes estudios relacionados con la evaluación del desarrollo sostenible, se identificaron los indicadores para cada dimensión; su elección se fundamentó en el marco de análisis definido por las naciones unidas, los ODS, y el resultado de un trabajo conjunto con actores del territorio y profesionales expertos en las temáticas inherentes a las dimensiones de la sostenibilidad. Además, se incluyó un análisis de información socializada a la comunidad en dichas temáticas, así como sus requerimientos.

A partir del conjunto de 81 indicadores, previamente seleccionados fue posible evaluar el nivel de avance en el desarrollo sostenible para cada año en el periodo 2009 a 2017, realizando además un análisis de estadística descriptiva y de correlación canónica para cada dimensión. La localidad registra en términos generales un nivel medio de sostenibilidad,

donde sólo los dos primeros años del análisis fueron evaluados en la categoría bajo. Además, la dimensión ambiental ha sido la más rezagada en comparación con las demás dimensiones. La evaluación realizada generó los insumos necesarios para el pronóstico de los niveles de sostenibilidad con herramientas de aprendizaje automático.

Es de resaltar que previo al desarrollo de un ejercicio de pronóstico, se identificaron y analizaron herramientas de aprendizaje automático, utilizadas en diferentes estudios relacionados con el desarrollo sostenible y en materia de calidad del aire. Para el periodo 2000 al 2019 se identificaron en las diferentes investigaciones las herramientas más utilizadas, así como los factores de éxito y aplicación; se generó un consolidado de las principales métricas de evaluación de desempeño, así como de las características requeridas en la operación del aprendizaje automático. Este análisis sentó las bases para su elección y, bajo la guía de las experiencias reportadas en las diferentes investigaciones, realizar su implementación.

Posteriormente y a partir del comportamiento identificado en la zona de estudio y con las herramientas seleccionadas, se desarrollaron modelos de clasificación de los niveles de desarrollo sostenible. Se realizó un pronóstico considerando la escala de información temporal, encontrando que las redes neuronales artificiales y los árboles de decisión presentaron las mejores métricas de desempeño; además, las variables de mayor influencia como el PM_{10} y el $PM_{2.5}$ se incluyeron como insumo en un posterior pronóstico de los niveles de desarrollo sostenible a nivel espacial. Este nuevo modelo permitió resolver una de las más importantes limitaciones que se presenta en estudios con la aplicación de herramientas de aprendizaje automático y que corresponde a la disponibilidad de información.

El modelo de predicción espacial del nivel de desarrollo sostenible y la consecuente identificación de la influencia de la calidad del aire en dicha clasificación, se llevó a cabo a partir de la composición de bandas de información espacial de 26 indicadores; fue posible observar en mayor detalle el comportamiento del micro territorio, en esta oportunidad en cada uno de los espacios que hacen parte de la localidad en estudio. En este modelo las herramientas con el mejor desempeño fueron las redes neuronales artificiales (96.5% y 88.2% para la clasificación de los niveles alto y medio), seguido del modelo de bosques aleatorios (96.8% para la clasificación del nivel bajo). Además, se encontró que el material particulado, en sus fracciones PM_{10} y $PM_{2.5}$ representó una importancia superior al 80% en el modelo con redes neuronales y al 40% en el modelo de clasificación con bosques aleatorios.

Finalmente, a partir de la realización de las actividades inherentes a cada objetivo específico se construyó una metodología para evaluar la influencia de la calidad del aire en el desarrollo urbano sostenible mediante el uso de herramientas de aprendizaje automático. Esta metodología es aplicable a zonas urbanas y orienta el paso a paso para la determinación de los factores de mayor relevancia en cada una de las dimensiones de la sostenibilidad, constituyéndose en un instrumento de soporte para la toma de decisiones respecto a la implementación y avance de ODS desde los micro territorios.

Abstract

Air quality is a determinant to the health and well-being of populations which, according to World Health Organization figures, has resulted in nearly seven million human lives lost each year. Furthermore, it is a key factor for the advancement of nations and territories at various levels, as human beings are the main focus of decisions on sustainable development. In this regard, within the framework of global public policy, a series of key elements have been designed and implemented, which are intended to overcome the challenges imposed by population growth and its consequent requirements. These include fundamental milestones such as the Millennium Development Goals and currently the Sustainable Development Goals (SDGs) as part of the 2030 Agenda. Additionally, these milestones have been supported by agreements, pacts, protocols, plans, programs and projects, all of which have been implemented at the country level through guidelines from the local level on up.

There are national progress reports on reaching specific goals, based on each country's agenda. In certain cases, these include city-level reports, whose indicators, both at the national and city levels, can be integrated into the central and best-known dimensions of sustainable development, namely the environmental, social, and economic dimensions. It is important to note that there is information concerning the monitoring of the state of resource quality and specific territorial conditions in various areas. However, not all territories in their different spatial scales are continuously evaluated for their sustainable performance. Moreover, environmental deterioration factors such as air pollution are handled as isolated determinants with reports generated on their behavior, in addition to developing monitoring and mitigation plans.

There are modelling tools to evaluate the behavior of different components that are part of the dimensions of sustainability. However, there is no instrument that forecasts the level of progress in sustainable development that also identifies the influence of air quality on its behavior. Machine learning tools can contribute to responding to this situation, as they are able to predict the behavior of a data set. These tools are based on regression and/or classification objectives, and through their results, support decision-making.

Therefore, the primary objective of this doctoral work is to establish the incidence of air quality on urban sustainable development, in its environmental, social, and economic dimensions, through the use of machine learning tools to support decision-making. This objective entails designing and implementing a methodology to identify the influence of air quality indicators on urban sustainable development. To this end, five specific objectives have been established, which consider the relevance of micro-territories as fundamental geographic settings to achieve the SDGs.

This doctoral work was performed based on a case study of a locality in the capital of Colombia; Bogotá, which covers an area of 350 km^2 and is located on an altitudinal plain in the eastern mountain range at 2625 meters above sea level. With 8.3 million inhabitants, Bogotá is one of the most populated cities in Latin America and is one of the world capitals

with the highest levels of air pollution from particulate matter, which is a risk factor for its population.

The following are the specific objectives proposed for the development of this research:

- Review the state of the art for identifying variables and parameters that could qualify the individual dimensions of sustainable performance as a function of the atmospheric component and air quality, as well as the time scale for the analysis.
- Perform statistical analysis of sustainable performance based on the variables and parameters selected for a specific locality in the city of Bogotá.
- Identify and select applicable machine learning tools.
- Use the selected machine learning tools to determine the incidence of air quality on urban sustainable development in a locality in the city of Bogotá.
- Obtain conclusions from the analysis and propose future developments.

This study began with gathering information on the locality in terms of air quality and health, and identifying the state of the art with respect to these factors. The behavior of atmospheric pollutants and meteorological variables was analyzed for the period 2009 – 2017. Furthermore, an analysis of the behavior of respiratory system diseases was developed, which was based on reports from government entities and other factors that include the proximity of homes to primary and secondary roads, as well as the results from fieldwork conducted in 2017. By using machine learning tools such as random forests and AdaBoost, and input information in raster formats structured from a composition of bands with the ArcGIS program and the free access software R, the areas of the locality with the highest probability of having respiratory system diseases were identified, as well as the main variables that determined this condition in the study area.

Afterward, the indicators were identified for each dimension based on a review and analysis of different studies related to the assessment of sustainable development. The framework analysis defined by the United Nations, the SDGs, and the results of joint work with actors in the territory and professional experts on the issues inherent to the sustainability dimensions supported the indicators' selection. Moreover, an information analysis on these subjects was included, which was published by online media. At the same time, an analysis was performed on community members' complaints and requests with respect to the Sustainable Development Goals.

The progress level in terms of sustainable development for each year from 2009 to 2017 was evaluated based on a set of 81 previously-selected indicators. Additionally, an analysis of descriptive statistics and canonical correlation for each dimension was performed. In general terms, the locality registered a medium level of sustainability, in which only the evaluation results for the first two years of the analysis scored into the "low" category. Moreover, the environmental dimension has lagged behind the other dimensions. The evaluation carried out generated the necessary inputs for forecasting sustainability levels with machine learning tools.

It is important to note that prior to the development of a prediction exercise, the machine learning tools used in different studies related to sustainable development and air quality were identified and analyzed. For the period 2000-2019, the most widely used tools in different research studies were identified, as well as their success and application factors. The primary performance evaluation metrics for machine learning were consolidated, as were the characteristics required for their operation. This analysis established the basis for their selection and based on the experiences reported in different research studies, their implementation.

Subsequently, and based on the behavior identified in the study area with the selected tools, classification models were developed for the levels of sustainable development. A forecast considering the temporal information scale was developed, which concluded that artificial neuronal networks and decision trees had the best performance metrics. Moreover, the most influential variables, namely PM_{10} and $PM_{2.5}$, were included as inputs in a subsequent forecast of sustainable development levels at the spatial level. This new model resolved one of the most important limitations presented in studies regarding the use of machine learning tools; information availability.

The development of a model for the spatial prediction of sustainable development levels and the consequent identification of the influence of air quality on this classification was performed based on the composition of information bands from 26 indicators. This enabled the examination of the micro-territory's behavior in greater detail; in this case, analyzing each space that makes up the locality under study. In this model, the tools with the best performance were artificial neural networks (96.5% and 88.2% for high- and medium-level classification, respectively), followed by the random forest model (96.8% for the low-level classification). Furthermore, it found that particulate matter, in fractions of PM_{10} and $PM_{2.5}$, represented importance greater than 80% in the neural network model and 40% in the random forest classification model.

Lastly, by carrying out the activities inherent to each specific objective, a methodology was developed to evaluate the influence of air quality on urban sustainable development with machine learning tools. This methodology is valid in urban areas, and through a step-by-step approach, determines the most relevant factors for each sustainability dimension. It has become a tool to support decision-making regarding the implementation and progress of the SDGs from the micro-territory level.

Resum

La qualitat de l'aire és un determinant de la salut i benestar de les poblacions que, d'acord amb xifres de l'Organització Mundial de la Salut, ha cobrat anualment prop de set milions de vides humanes. És a més un factor clau per a l'avanç de les nacions i dels territoris en les seues diferents escales, perquè és l'ésser humà l'eix principal de les decisions en matèria de desenvolupament sostenible. Sobre aquest tema, en el marc de la política pública a nivell mundial, s'han dissenyat i posat en marxa una sèrie d'elements clau, la fi dels quals ha sigut superar els desafiaments que imposa el creixement poblacional i els seus conseqüents requeriments; es tracta de grans fites com els Objectius de Desenvolupament del Mil·lenni i en l'actualitat, els Objectius de Desenvolupament Sostenible (*ODS) amb l'Agenda 2030. Aquests elements s'han vingut implementant amb el suport d'acords, convenis, protocols, plans, programes i projectes que des de l'àmbit de país són adreçats i decantats per a la seua implementació.

Existeixen reportes nacionals d'avanç sobre la implementació de metes específiques, segons l'agenda de cada país i en alguns casos en l'àmbit de ciutat, els indicadors de la qual poden integrar-se en les dimensions centrals i més conegudes del desenvolupament sostenible: la dimensió ambiental, la social i l'econòmica. Es destaca que existeix informació sobre el monitoratge de l'estat de la qualitat dels recursos i de les condicions específiques del territori en diversos temes. No obstant això, no en tots els territoris, en les seues diferents escales espacials, es realitza contínua avaluació del seu acompliment sostenible i, a més factors de deterioració ambiental com la contaminació de l'aire, són tractats com a determinants aïllats amb la generació de reportes del seu comportament i el desenvolupament de plans de monitoratge i de mitigació.

Per als diferents temes que fan part de les dimensions de la sostenibilitat, existeixen eines de modelatge per a avaluar el comportament dels seus indicadors; no obstant això, no es compta amb un instrument que pronostique el nivell d'avanç en el desenvolupament sostenible i a més que identifique la influència de la qualitat de l'aire en el seu comportament. Les eines d'aprenentatge automàtic poden aportar en la resposta a aquesta situació, atés que consisteixen en una estructura útil per al pronòstic del comportament d'un conjunt de dades, a partir d'uns objectius de regressió i/o classificació i que suporten amb els seus resultats a la presa de decisions.

Per consegüent, l'objectiu central d'aquest treball doctoral és establir la incidència de la qualitat de l'aire sobre el desenvolupament urbà sostenible, en les seues dimensions ambiental, social i econòmica, mitjançant l'ús d'eines d'aprenentatge automàtic, com a suport per a la presa de decisions. Aquest objectiu involucra el disseny i execució d'una metodologia per a identificar la influència d'indicadors en matèria de qualitat de l'aire, sobre el desenvolupament urbà sostenible. Per al seu assoliment s'han definit cinc objectius específics, que consideren la rellevància dels micro territoris com a escenari geogràfic fonamental a l'assoliment dels ODS.

Aquest treball doctoral es duu a terme a partir d'un estudi de cas en una localitat de la ciutat de Bogotà, a Colòmbia que és la capital del país, amb una extensió de 350 km², situada sobre una planícia altitudinal en la serralada oriental i a 2625 metres sobre el nivell de la mar. Bogotà és una de les ciutats més poblades a Amèrica Llatina amb 8.3 milions d'habitants i és una de les capitals mundials que ha presentat alts nivells de contaminació per material particulat, sent aquest un factor de risc per a la seua població.

Els següents són els objectius específics plantejats pel desenvolupament d'aquesta investigació:

- Revisar l'estat de l'art per a la identificació de les variables i paràmetres que podrien qualificar les dimensions individuals de l'acompliment sostenible en funció del component atmosfèric i qualitat de l'aire, així com l'escala temporal per a l'anàlisi
- Realitzar una anàlisi estadística d'acompliment sostenible d'acord amb les variables i paràmetres seleccionats per a una localitat específica a la ciutat de Bogotà
- Identificar i seleccionar les eines d'aprenentatge automàtic aplicables
- Aplicar les eines d'aprenentatge automàtic seleccionades per a determinar la incidència de la qualitat de l'aire sobre el desenvolupament urbà sostenible en una localitat de la ciutat de Bogotà
- Obtindre conclusions sobre l'anàlisi i proposar desenvolupaments futurs

Aquest treball es va iniciar amb el coneixement de localitat de Kennedy en termes de qualitat de l'aire i salut, identificant l'estat de l'art en aquests factors; es va analitzar el comportament de contaminants atmosfèrics i variables meteorològiques en el període 2009 a 2017, el comportament de malalties del sistema respiratori, segons reportes entregats per entitats governamentals, i altres factors que inclouen la proximitat de les llars a vies primàries i secundàries, a més dels resultats d'un treball de camp realitzat en 2017. Amb l'aplicació de les eines d'aprenentatge automàtic denominades Random Forest i Adaptative boosting, així com informació d'entrada format ráster, estructurada a partir d'una composició de bandes amb l'ús del programari ArcGis i el llenguatge de programació d'accés lliure R, es van determinar les zones de la localitat amb major probabilitat de presentar malalties del sistema respiratori, alhora que les principals variables que van determinar aquesta condició en el territori.

Posteriorment, a partir d'una revisió i anàlisi de diferents estudis relacionats amb l'avaluació del desenvolupament sostenible, es van identificar els indicadors per a cada dimensió; la seua elecció es va fonamentar en el marc d'anàlisi definida per les nacions unides, els *ODS, i el resultat d'un treball conjunt amb actors del territori i professionals experts en les temàtiques inherents a les dimensions de la sostenibilitat. A més, es va incloure una anàlisi d'informació socialitzada a la comunitat en aquestes temàtiques, així com els seus requeriments.

A partir del conjunt de 81 indicadors, prèviament seleccionats va ser possible avaluar el nivell d'avanç en el desenvolupament sostenible per a cada any en el període 2009 a 2017, realitzant a més una anàlisi d'estadística descriptiva i de correlació canònica per a cada dimensió. La localitat registra en termes generals un nivell mitjà de sostenibilitat, on només els dos primers

anys de l'anàlisi van ser avaluats en la categoria baix. A més, la dimensió ambiental ha sigut la més ressagada en comparació amb les altres dimensions. L'avaluació realitzada va generar els inputs necessaris pel pronòstic dels nivells de sostenibilitat amb eines d'aprenentatge automàtic.

És de ressaltar que previ al desenvolupament d'un exercici de pronòstic, es van identificar i van analitzar eines d'aprenentatge automàtic, utilitzades en diferents estudis relacionats amb el desenvolupament sostenible i en matèria de qualitat de l'aire. Per al període 2000 al 2019 es van identificar en les diferents investigacions les eines més utilitzades, així com els factors d'èxit i aplicació; es va generar un consolidat de les principals mètriques d'avaluació d'acompliment, així com de les característiques requerides en l'operació de l'aprenentatge automàtic. Aquesta anàlisi va establir les bases per a la seua elecció i, sota la guia de les experiències reportades en les diferents investigacions, realitzar la seua implementació.

Posteriorment i a partir del comportament identificat en la zona d'estudi i amb les eines seleccionades, es van desenvolupar models de classificació dels nivells de desenvolupament sostenible. Es va realitzar un pronòstic considerant l'escala d'informació temporal, trobant que les xarxes neuronals artificials i els arbres de decisió van presentar les millors mètriques d'acompliment; a més, les variables de major influència com el PM_{10} i el $PM_{2.5}$ es van incloure com a input en un posterior pronòstic dels nivells de desenvolupament sostenible a nivell espacial. Aquest nou model va permetre resoldre una de les més importants limitacions que es presenta en estudis amb l'aplicació d'eines d'aprenentatge automàtic i que correspon a la disponibilitat d'informació.

El model de predicció espacial del nivell de desenvolupament sostenible i la conseqüent identificació de la influència de la qualitat de l'aire en aquesta classificació, es va dur a terme a partir de la composició de bandes d'informació espacial de 26 indicadors; va ser possible observar en major detall el comportament del micro territori, en aquesta oportunitat en cadascun dels espais que fan part de la localitat en estudi. En aquest model les eines amb el millor acompliment van ser les xarxes neuronals artificials (96.5% i 88.2% per a la classificació dels nivells alt i mitjà), seguit del model de boscos aleatoris (96.8% per a la classificació del nivell baix). A més, es va trobar que el material particulat, en les seues fraccions PM_{10} i $PM_{2.5}$ va representar una importància superior al 80% en el model amb xarxes neuronals i al 40% en el model de classificació amb boscos aleatoris.

Finalment, a partir de la realització de les activitats inherents a cada objectiu específic es va construir una metodologia per a avaluar la influència de la qualitat de l'aire en el desenvolupament urbà sostenible mitjançant l'ús d'eines d'aprenentatge automàtic. Aquesta metodologia és aplicable a zones urbanes i orienta el pas a pas per a la determinació dels factors de major rellevància en cadascuna de les dimensions de la sostenibilitat, constituint-se en un instrument de suport per a la presa de decisions respecte a la implementació i avanç dels ODS des dels micro territoris.

Contenido

Índice de Figuras

Índice de Tablas

Índice de Abreviaturas

AHP	Análisis de priorización jerárquico
ANN	Redes neuronales artificiales
AUC	Área bajo la curva
CO	Monóxido de carbono
DUS	Desarrollo urbano sostenible
DT	Árboles de decisión
IDW	Interpolación de distancia inversa ponderada
IPM	Índice de pobreza multidimensional
K-nn	K vecinos más próximos
NDVI	Índices diferencial de vegetación normalizada
NO_X	Óxidos de nitrógeno
ODS	Objetivos de desarrollo sostenible
OMS	Organización Mundial de la Salud
PM_{10}	Material particulado en la fracción inferior a 10 micras
$PM_{2.5}$	Material particulado en la fracción inferior a 2.5 micras
RF	Bosques aleatorios
ROC	Característica operativa del receptor
SVM	Máquina de soporte vectorial
SOx	Óxidos de Azufre
T	Temperatura
TST	Temperatura superficial terrestre
UPZ	Unidad de Planeación Zonal

1 Capítulo 1. Introducción

Las ciudades forman parte de un área geográfica en donde interactúan espacios construidos y zonas amortiguadoras garantes de servicios ecosistémicos, los cuales están influenciados por actividades económicas de orden manufacturero, de prestación de servicios, de extracción de minerales y productos de la tierra, propios para garantizar un nivel de vida acorde a un estándar definido en la población que las habita. Al ser espacios abastecedores de servicios se enfrentan a grandes retos: seguridad, salud, contaminación, construcciones y asentamientos ilegales, población e infraestructuras ubicadas en zonas vulnerables, ausencia de zonas verdes, cobertura de servicios básicos, conectividad del transporte, pobreza, inequidad, violencia, alimentación, educación, trabajo, expansión en el uso del suelo urbano, vivienda digna, acceso a los servicios de salud y calidad ambiental, entre otros.

Las condiciones económicas, políticas y las dificultades propias de los territorios orientan procesos de migración continuos hacia las zonas urbanas, incrementando la presión sobre los recursos y exigiendo de éstas la satisfacción de las necesidades y oferta de soluciones a los problemas que plantea el incremento poblacional. De acuerdo con cifras del Banco Mundial cerca del 55% de la población vive en zonas urbanas (Banco Mundial, 2020), lo cual influye en su extensión y las presiones sobre los servicios ecosistémicos que esto conlleva. El deterioro de la calidad del aire es uno de los grandes retos a nivel urbano, además de la disponibilidad de agua potable y condiciones apropiadas de saneamiento básico; estos factores influencian la salud y el bienestar de la población.

El desarrollo urbano sostenible (DUS) es un reto que incluye superar los desafíos establecidos en los territorios; esto supone la interacción entre los objetivos, metas y acciones para alcanzar un desarrollo a partir de la integración de las dimensiones ambiental, social y económica bajo el soporte de la articulación interinstitucional. La dimensión ambiental ofrece las condiciones esenciales para la vida, la dimensión social constituye los objetivos deseables de bienestar y calidad de vida del ser humano, y la dimensión económica incluye los medios para alcanzar dichos objetivos (Boivin and Tanguay, 2018). Se trata de un desarrollo que garantice espacios habitables con miras a un bienestar social, mejor calidad ambiental y de vida de la población; un desarrollo viable orientado a un crecimiento económico, consciente de la capacidad de carga de los ecosistemas y, un desarrollo equitativo que incluya el respeto por la dignidad del ser humano y el equilibrio en el acceso a los bienes y servicios que se ofertan en el área urbana.

Con el fin de evaluar el nivel de avance en la sostenibilidad de países y/o ciudades se han realizado diversos estudios (Antanasijević et al., 2017; Mirshojaeian y Kaneko, 2011; Nilashi et al., 2019; Pérez-Ortíz et al., 2014; Toumi et al., 2017) a partir del análisis de temas, subtemas e indicadores específicos de los pilares de la sostenibilidad, en diferentes periodos de tiempo o para un momento determinado. Se han implementado metodologías de evaluación que parten de la definición de un marco de análisis y en su mayoría corresponden a un enfoque de recopilación de información temporal y su análisis individual en periodos de tiempo establecidos. Pocos estudios han indagado en el comportamiento específico de la

sostenibilidad urbana a nivel espacial, incluidos el análisis del comportamiento de las dimensiones de la sostenibilidad (Shen et al., 2013).

En la evaluación de la sostenibilidad se encuentra como una de sus limitantes, la disponibilidad de información para cada uno de los indicadores objeto de análisis; es por ello que, la mayoría de los estudios contemplan evaluaciones a nivel de país, a nivel de ciudad e incluso a nivel sectorial y empresarial, dejando de lado al ámbito local más específico de las ciudades correspondiente a los micro territorios. Además, la identificación de los indicadores a nivel territorial obedece a un ejercicio liderado principalmente por entidades gubernamentales con una visión específica de objetivos, metas, planes a ejecutar y, en algunos casos, ausente de la integración de las necesidades específicas de la población en el territorio.

Se destaca que el alcance de las metas y objetivos establecidos a partir de la Agenda 2030, en el marco de los objetivos de desarrollo sostenible (ODS) (Organización de las Naciones Unidas, 2018), requiere del conocimiento de los territorios en todas sus escalas, en donde el efecto sinérgico de sus acciones se suma a la implementación de medidas en el ámbito nacional. Para ello, es necesario contar con un conjunto de información y del uso de herramientas apropiadas para su captura, tratamiento y análisis. Llama la atención la aplicación de modelos basados en el aprendizaje automático, a partir de la definición de unas condiciones de entrenamiento de un set de datos, donde la herramienta identifica patrones de comportamiento y genera escenarios futuros como base de información para la definición de estrategias de gestión.

El tema central de este trabajo de doctorado corresponde con el diseño de una metodología para la identificación de la influencia de la calidad del aire en el desarrollo urbano sostenible, a partir de herramientas de aprendizaje automático. Su desarrollo analiza las limitantes de información y busca estructurar una metodología que, no sólo establezca un análisis temporal, sino además la identificación del comportamiento espacial territorial. De acuerdo con (Whitehead, 2003) las ciudades sostenibles son espacios constituidos de manera individual en escalas temporales y geográficas específicas, por lo que este trabajo parte de la necesidad de analizar a los micro territorios o zonas que conforman a las ciudades y que representan un valor administrativo, económico, ambiental y político importante para los territorios urbanos y la población que allí habita.

El análisis de la influencia de la calidad del aire en el desempeño urbano sostenible permite interpretar la magnitud de su relevancia y su relación con las condiciones del territorio que propician su estado. Diversas investigaciones analizan su comportamiento individual y su influencia en la salud de la población, indicando los efectos en el estado de la calidad del aire y en la salud dada la cercanía a fuentes de emisión, incluidas las fuentes móviles y fuentes fijas; estos estudios han analizado la manera en la que contaminantes como el material particulado en sus fracciones menores a 10 y 2.5 micras (PM_{10} y $PM_{2.5}$, respectivamente) contribuyen con la incidencia y prevalencia de enfermedades cardiovasculares, respiratorias y enfermedades mentales (Lary et al., 2015; Li et al., 2011; Tajudin et al., 2019). Sin embargo, son limitados los estudios que analizan la articulación de las condiciones de calidad

del aire con su intervención en el desempeño sostenible territorial; al ser uno de los problemas urbanos de mayor relevancia, requiere de un análisis específico en lo asociado a la sostenibilidad.

Las limitaciones a las que se enfrentó este estudio corresponden con la disponibilidad de información, debidas principalmente al alcance territorial priorizado y escasez de bases de datos centralizadas con la información específica del territorio; se destaca que sólo a nivel nacional y en el ámbito de ciudad existe mayor disponibilidad de información. Por lo tanto, se hizo uso de un conjunto de herramientas de selección, recopilación y análisis de la información y posteriormente, se aplicaron modelos en los que las bondades del aprendizaje automático permitieron identificar escenarios de comportamiento del set de datos analizado.

Las herramientas de aprendizaje automático son instrumentos cuya operatividad y recursos tecnológicos requeridos, las hacen útiles en la resolución de problemas como los planteados en este trabajo. La articulación de estas herramientas con los procedimientos avanzados en el marco global, para la evaluación del desempeño sostenible, establece una combinación apropiada para la definición de elementos específicos en los micro territorios para la optimización de acciones, medidas y presupuestos.

Se espera que el contenido de este documento, producto del trabajo de doctorado, sea un insumo en la planificación de ciudades desde el nivel de micro territorios y en el soporte en la toma de decisiones, desde la recopilación de información, como en la definición de estrategias, con miras a avanzar hacia un desarrollo urbano sostenible.

El contenido de este documento está distribuido de la siguiente manera: los dos primeros capítulos aportan una breve introducción en la temática desarrollada, incluido un breve análisis del estado del arte y la presentación, justificación y análisis de los objetivos de este trabajo. Posteriormente, se presenta un capítulo con los materiales y métodos utilizados en este trabajo, seguido del capítulo de resultados y discusión; finaliza este documento de tesis doctoral con el capítulo de conclusiones y desarrollos futuros, dando cumplimiento al quinto y último objetivo específico.

La metodología para determinar la influencia de la calidad del aire en el desarrollo urbano sostenible, se construyó a través de los objetivos que se presentan en el segundo capítulo de este documento; se diseñó la metodología con el apoyo de un estudio de caso, reconociendo las dificultades asociadas a la disponibilidad de información y la importancia del análisis de micro territorios como soporte al avance en la sostenibilidad de las naciones, en el mejoramiento de la calidad de vida y en salud de la población, en el contexto de la influencia ejercida por la calidad del aire.

Este trabajo está basado en la colección de publicaciones, que pueden consultarse en los apéndices A, B, C, D y E del presente documento a propósito de profundizar en los aspectos desarrollados en los cinco capítulos de este trabajo.

2 Capítulo 2. Objetivos y Justificación

La contaminación atmosférica ha sido un aspecto estudiado en diferentes escenarios, dada la manifestación de eventos con altas tasas de morbilidad y mortalidad en poblaciones expuestas de manera aguda o crónica a contaminantes. Es uno de los impactos más relevantes en los contextos urbanos, que supone la presencia de fuentes de emisión industriales, vehiculares, la escasez de zonas amortiguadoras de aspectos ambientales y aquellas relacionadas con el deterioro de edificaciones y vías. Su influencia en la calidad del aire ha sido protagonista de documentos científicos que dan cuenta de las herramientas para su monitoreo, para el control y eliminación de las emisiones de contaminantes atmosféricos, además de aquellas herramientas que buscan establecer el grado de relación de dichos contaminantes con los efectos en la salud de la población a corto, mediano y largo plazo.

De acuerdo con la Organización Mundial de la Salud (OMS) cada año mueren 7 millones de personas a causa de la contaminación del aire (OMS, 2018); dada su relevancia se ha incluido como parte de los ODS con el fin de reducir el impacto ambiental negativo per cápita de las ciudades (Organización de las Naciones Unidas, 2018). Sin embargo, aun cuando el ser humano es el eje central de los retos del desarrollo sostenible, no se ha identificado un estudio que analice la influencia de la calidad del aire en dicho desarrollo; por lo que el objetivo central de este trabajo doctoral es establecer la incidencia de la calidad del aire sobre el desarrollo urbano sostenible, en sus dimensiones ambiental, social y económica, mediante el uso de herramientas de aprendizaje automático, como soporte para la toma de decisiones. Este objetivo involucra el desarrollo de una metodología para identificar la influencia de indicadores, en materia de calidad del aire, sobre el desarrollo urbano sostenible.

Para el logro de dicho objetivo se han establecido cinco objetivos específicos, en donde el primero de estos correspondió con: revisar el estado del arte para la identificación de las variables y parámetros que podrían calificar las dimensiones individuales del desempeño sostenible en función del componente atmosférico y calidad del aire, así como la escala temporal para el análisis; el segundo objetivo específico consistió en realizar un análisis estadístico del desempeño sostenible, de acuerdo con las variables y parámetros seleccionados para una localidad específica en la ciudad de Bogotá; seguido del tercer objetivo que correspondió con: identificar y seleccionar las herramientas de aprendizaje automático aplicables; el cuarto objetivo orientado a aplicar las herramientas de aprendizaje automático seleccionadas para determinar la incidencia de la calidad del aire sobre el desarrollo urbano sostenible en una localidad específica de la ciudad de Bogotá. Finalmente, el quinto y último objetivo específico, consistente en obtener conclusiones acerca del análisis y proponer desarrollos futuros, el cual se realizó de manera paralela a cada uno de los objetivos previos.

Un análisis de la incidencia de la calidad del aire sobre el desarrollo urbano sostenible requiere conocer las variables y parámetros que lo califican y a sus dimensiones individuales, en especial en los contextos urbanos y en su relación con los fines de los ODS; del mismo modo, es preciso comprender la escala temporal para el análisis, dichos elementos fueron identificados a partir del desarrollo del primer objetivo específico.

Al respecto, diversos estudios han analizado la sostenibilidad de los territorios aplicando diferentes metodologías que incluyen, entre otras, el barómetro de sostenibilidad, el tablero de sostenibilidad y el cálculo del índice de sostenibilidad bajo la aplicación de análisis multicriterio, opinión de expertos, así como el análisis de componentes principales (Antanasijević et al., 2017; Carrillo-Rodríguez y Toca, 2013; Meijering et al., 2018; Mirshojaeian y Kaneko, 2011; Prescott-allen, 1997; Scipioni et al., 2009; Singh et al., 2012; Toumi et al., 2017). La mayoría de los estudios han sido aplicados a grandes ciudades con una escala temporal anual, dejando de lado el análisis de territorios a una menor escala territorial (micro territorios), quienes en últimas también tienen como objetivo el cumplimiento de los fines de la Agenda 2030 y, además, aportan desde el contexto urbano no sólo, a las condiciones de la calidad del aire, sino también a los efectos en la salud desde el nivel del micro territorio al contexto nacional.

Para conocer el comportamiento de los micro territorios se requiere de la apropiada elección de parámetros e indicadores, cuyas características permitan aportar con la información suficiente para el análisis de la sostenibilidad y sus dimensiones. De allí que, el desarrollo del primer objetivo específico, aporte como un primer eslabón en la formulación de la metodología de análisis.

Del mismo modo, el análisis de los indicadores elegidos permitirá comprender el comportamiento temporal de las variables estudiadas, por lo que la aplicación de técnicas estadísticas aporta al conocimiento en el comportamiento de dichos indicadores dentro de la dimensión a la cual pertenecen (ambiental, social, económica, institucional), en la relación con indicadores de otras dimensiones y, en el comportamiento del desempeño del territorio analizado en el marco del desarrollo sostenible. En este orden de ideas, el análisis estadístico del desempeño sostenible, de acuerdo con las variables y parámetros seleccionados, como segundo objetivo específico, permitió identificar el comportamiento de las variables, sus interrelaciones y el grado de dependencia entre ellas, así como el comportamiento de los niveles de sostenibilidad en cada periodo de estudio.

Como se ha mencionado, el cálculo del nivel de sostenibilidad ha sido abordado en diferentes trabajos a través de la aplicación de indicadores a escala anual, estableciendo para determinado periodo un nivel de avance en la sostenibilidad. Sin embargo, este cálculo se constituye en una fotografía de un momento específico, ya que no permite identificar o prever el comportamiento de las dimensiones, sus indicadores y el desempeño sostenible del territorio como un todo.

Además, no se ha desarrollado un estudio en el que se identifique y pronostique el comportamiento de las dimensiones de la sostenibilidad y sus indicadores a nivel espacial en el micro territorio, por lo que éste es el primero de los estudios desarrollado en la materia y cuyos resultados a través de la definición de una metodología, se consolidan en una herramienta necesaria en la priorización de medidas y recursos para el avance en la sostenibilidad urbana, incluidos temas clave como la calidad del aire.

Por lo tanto, el uso de herramientas de aprendizaje automático permitió pronosticar el comportamiento de los niveles de sostenibilidad, así como establecer el grado de importancia

de las variables de entrada para identificar la influencia de la calidad del aire en el desarrollo urbano sostenible. Para ello, fue necesario el desarrollo del tercer objetivo específico consistente en la identificación y selección de las herramientas de aprendizaje automático aplicables.

Una vez identificadas y seleccionadas las herramientas de aprendizaje de máquina, se procedió a pronosticar el nivel de desempeño sostenible, tanto en una escala temporal como en el contexto espacial, dando alcance al cuarto objetivo específico consistente en: aplicar las herramientas de aprendizaje automático seleccionadas para determinar la incidencia de la calidad del aire sobre el desarrollo urbano sostenible en una localidad de la ciudad de Bogotá. El grado de influencia de las variables en el proceso de clasificación se determinó a partir de los modelos de pronóstico de mejor desempeño.

3 Capítulo 3. Materiales y Métodos

La formulación de una metodología para la determinación de la incidencia de la calidad del aire sobre el desarrollo urbano sostenible (DUS), en sus dimensiones ambiental, social y económica, mediante la aplicación de herramientas de aprendizaje automático se centra en cinco aspectos principales: 1) la elección del estudio de caso y el conocimiento de las características de la zona elegida, que en el marco del primer objetivo específico, permitió sentar las bases para la identificación de las variables y parámetros que podrían calificar cada una de las dimensiones del desempeño sostenible, considerando el componente atmosférico, calidad del aire y la escala temporal para el análisis; 2) la evaluación del desempeño sostenible como soporte en la ejecución del segundo objetivo específico; 3) la identificación y definición de herramientas de aprendizaje automático, en respuesta al tercer objetivo específico planteado en este proyecto; 4) el pronóstico del desempeño sostenible a partir de la aplicación de las herramientas seleccionadas y, 5) la determinación de la influencia de la calidad del aire en el desarrollo urbano sostenible (ver figura 3-1), siendo esta la etapa final para lograr el cumplimiento del objetivo central de este trabajo doctoral.

Dichas etapas se desarrollaron a través de una serie de actividades que son descritas en los siguientes apartados y que soportan el diseño de la estructura metodológica definida a lo largo de esta investigación.

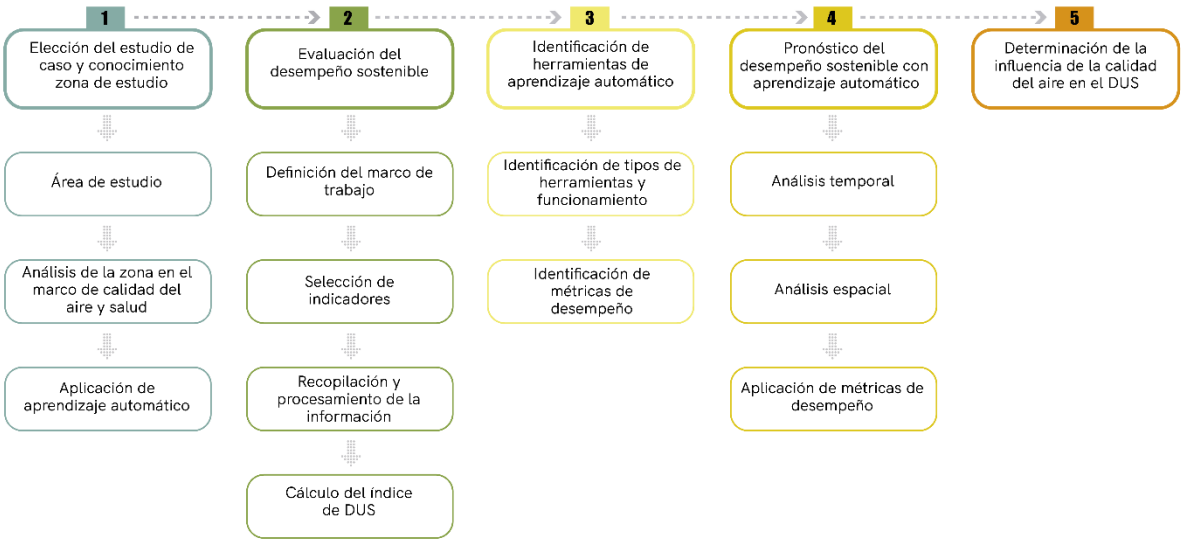


Figura 3-1 Esquema general para la identificación de la influencia de la calidad del aire en el desarrollo urbano sostenible

3.1 Elección del estudio de caso y conocimiento de las características de la zona elegida

El desarrollo sostenible es un concepto que ha orientado principalmente a las naciones en la implementación de diversas medidas para combatir la pobreza, el hambre, así como la inequidad. Al punto en que las naciones han suscrito acuerdos y compromisos para avanzar en la implementación de las metas definidas alrededor de los objetivos de desarrollo del

milenio (2000-2015) y por otra parte al alcance de los objetivos de desarrollo sostenible (ODS) con la Agenda 2030 (2016-2030). Sin embargo, para su logro es necesario identificar los aportes al cumplimiento de las metas fijadas en el orden nacional, desde los territorios de menor escala espacial y que hacen parte de las ciudades, lo cual involucra a los micro territorios. Por lo tanto, para establecer los aportes a las metas fijadas, cada territorio requiere de un análisis independiente dadas sus particularidades a nivel ambiental, social y económico.

Considerando los fines de este estudio se eligió un territorio urbano que presenta niveles característicos de contaminación del aire, alta densidad poblacional y el desarrollo de diversas actividades económicas; para ello se realizó un análisis comparativo de las 20 localidades de la ciudad de Bogotá, Colombia; seleccionando la localidad con los mayores niveles registrados de contaminación del aire en los últimos 5 años, diversidad de actividades económicas, así como alto número de habitantes en comparación con las demás localidades.

Las características del territorio seleccionado permitirán modelar el comportamiento común de zonas urbanas con sus problemas o situaciones independientes, pero que pueden ser representativas en el contexto urbano a nivel global. Se eligió a la localidad de Kennedy ubicada en la ciudad de Bogotá en Colombia.

3.1.1 Descripción del área de estudio

La localidad de Kennedy es una zona urbana ubicada al sur occidente de Bogotá, la capital de Colombia (ver figura 3-2) que a su vez está distribuida en 20 localidades. Kennedy presenta una extensión de 38.58 km^2 y una población cercana a los 1.2 millones de habitantes; esta población corresponde al 14.5% de la población de la ciudad capital. Esta localidad es una de las más densamente pobladas de Bogotá, que a su vez es una de las ciudades más densamente pobladas a nivel mundial.

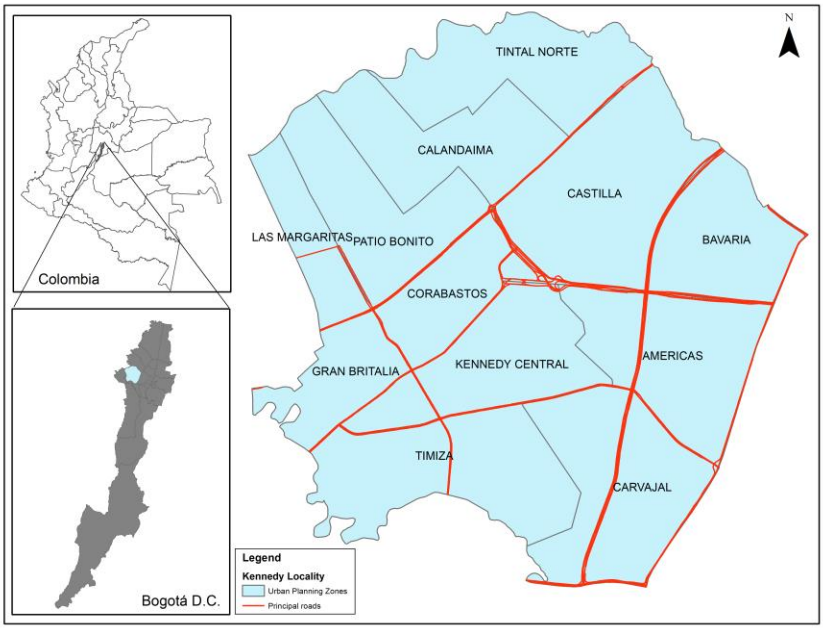


Figura 3-2 Localidad de Kennedy, su distribución administrativa, usos de suelo y vías principales (Molina-Gómez, et al., 2021a)

Kennedy concentra diversidad de actividades que interactúan entre los diferentes usos del suelo, dentro de los cuales se incluyen los usos residenciales, usos dotacionales, de comercio y servicios, y la presencia de actividades de orden industrial, distribuidos en doce unidades de planificación zonal (UPZ) (ver figura 3-2). Al ser las UPZ instrumentos que definen el planeamiento del suelo urbano, según la dinámica productiva de la ciudad y su inserción en el contexto regional (Secretaría Distrital de Planeación, 2020), son de gran relevancia en la organización y dinámica de la localidad. Al respecto, cinco UPZ de Kennedy son de uso exclusivamente residencial (Carvajal, Timiza, y Castilla); tres de uso no exclusivamente residencial (Patio Bonito, Gran Britalia y Corabastos); una en uso industrial (Bavaria) y las restantes en usos de centralidad urbana (Américas), desarrollo urbano (Calandaima y Tintal Norte) y utilidad pública (Calandaima).

Adicionalmente, al ser una localidad que integra a la principal central de almacenamiento y venta de productos alimenticios de la capital del país, está rodeada por vías de alto flujo vehicular y en su interior vías que permiten el ingreso y salida tanto de vehículos de carga como la entrada, salida y circulación de vehículos de transporte público de pasajeros y vehículos privados.

Kennedy es una localidad que ha presentado un proceso de urbanización con el consecuente incremento de la población en las doce diferentes UPZ que la constituyen (ver figura 3-2). Además del incremento en la presión sobre los recursos a partir de emisiones atmosféricas, residuos sólidos y descarga de vertimientos no controlados en fuentes hídricas, esta zona urbana ha revelado un incremento en la temperatura superficial terrestre y la limitación de los recursos amortiguadores de dichos aspectos ambientales (ver apéndice E).

3.1.2 Análisis de la zona de estudio en materia de calidad del aire y salud

Se realizó la caracterización de la zona y el análisis de la influencia de la calidad del aire en la salud de la población, específicamente en lo relacionado con enfermedades del sistema respiratorio, dado que las infecciones respiratorias, la enfermedad obstructiva crónica y el cáncer de pulmón son algunas de las principales consecuencias relacionadas con la exposición a altos niveles de contaminación del aire. A partir de ello, fue posible establecer un panorama del territorio analizado en relación con la presencia de posibles casos de enfermedad respiratoria en sus diferentes zonas, dada la variedad de actividades, usos de suelo, contaminantes atmosféricos y condiciones climáticas entre otros aspectos que pueden influir. Para ello se aplicó el procedimiento esquematizado en la figura 3-3 y que se describe a continuación:

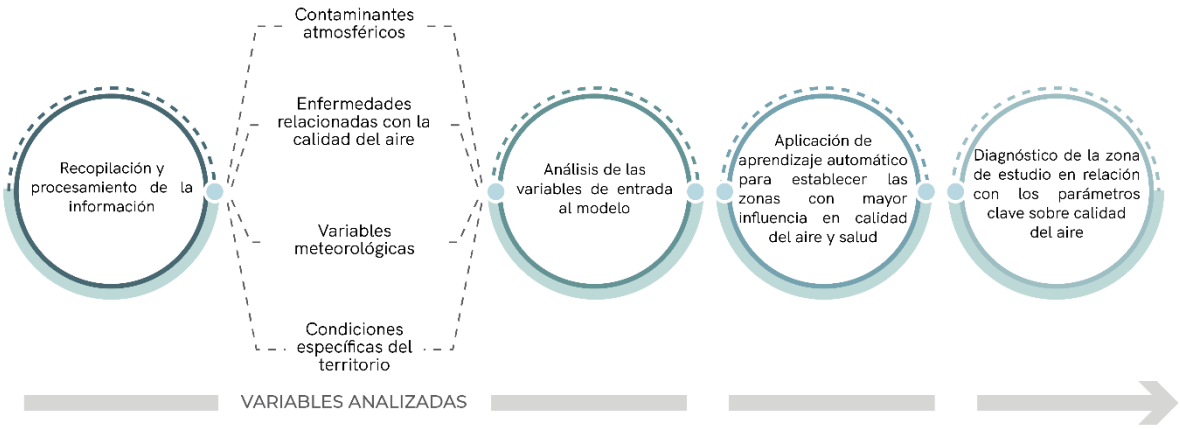


Figura 3-3 Procedimiento general para el análisis espacial de la relación calidad del aire y salud respiratoria

3.1.2.1 Recopilación y procesamiento de la información

Se realizó recopilación de la información de los parámetros (PM_{10}, $PM_{2.5}$, SO_X, NO_X, CO) y variables meteorológicas (precipitación, temperatura) para el periodo 2009-2017, y se eligió el 2016 como año base para el análisis espacial de comportamiento de la información, siendo el 2016 el año inmediatamente anterior al desarrollo de actividades en campo. Para cada una de las variables se identificó su comportamiento a partir de las siguientes métricas de estadística descriptiva: valores máximos, mínimos, valores promedio de los datos diarios y del comportamiento anual de cada variable.

Mediante la aplicación del método de interpolación de distancia inversa ponderada IDW, por sus siglas en inglés, se estableció la distribución espacial de los contaminantes y variables meteorológicas. Este método consiste en la predicción de un valor a partir de datos conocidos en un espacio geográfico, en donde dichos valores ejercerán la mayor influencia en cuanto más cerca se encuentre del punto de predicción; es decir, al existir mayor distancia entre el punto de datos y el punto a predecir, menor será el valor de ponderación asignado a la predicción (Mesnard, 2013)

Se aplicó este método de interpolación considerando la distribución no uniforme en la que se localizan las estaciones de monitoreo de calidad del aire en el área de influencia de la zona de estudio (estaciones: Tunal, Puente Aranda, Simón Bolívar, Kennedy, Carvajal en Bogotá y Mosquera-Sena en Cundinamarca). Para su aplicación se hizo uso de los datos de contaminantes y variables meteorológicas medidos en las estaciones de monitoreo, el parámetro de potencia p=2 que permitió controlar la significancia de los puntos conocidos en los valores interpolados, con un reducción de la importancia en función de la distancia al cuadrado (Gómez-Losada et al., 2019); y un radio de búsqueda de tipo variable, éste último permitió que en el método se considerara la distancia a cada una de las estaciones proveedoras de los datos. Adicionalmente, se abordó el método de clasificación de la información interpolada a partir del sistema propuesto por Jenks (1967), con el fin de optimizar los datos bajo características similares.

Se recopilaron estadísticas de morbilidad y mortalidad para el mismo periodo a partir de información reportada por la Secretaría Distrital de Salud, asociadas a enfermedades del

sistema respiratorio, las cuales se analizaron mediante métricas de estadística descriptiva. Considerando que por seguridad la información espacial asociada a los registros de morbilidades no es de acceso público, se llevó a cabo en 2017 un trabajo de campo en el que se identificaron eventos diagnosticados por un médico en 2016 correspondientes a enfermedad respiratoria o infección como asma, neumonía o enfermedad severa de los pulmones. Este trabajo consistió en la aplicación de un formato de recopilación de información con una muestra total de 912 hogares, teniendo en cuenta un nivel de confianza del 96% y un error de 0.04, establecidos a partir de un análisis de la población proyectada para cada una de las UPZ que conforman a la localidad de Kennedy. El detalle de este procedimiento puede consultarse en el apéndice A.

3.1.2.2 Análisis de las variables de entrada en el modelo

Para determinar el grado de dependencia de las variables de entrada al modelo se realizó un análisis de correlación de Pearson a partir de información tipo ráster. Las variables analizadas fueron las siguientes: contaminantes atmosféricos ($PM_{10,}$ $PM_{2.5}$, SO_2, NO_X, CO) y variables meteorológicas (precipitación, temperatura), población existente en cada UPZ, densidad poblacional según UPZ y uso del suelo, cuyas características se detallan en el apéndice A.

3.1.2.3 Zonas de interés en términos de calidad del aire y salud

La identificación de aquellas zonas en las que podrían presentarse casos de enfermedad respiratoria se realizó mediante el uso de herramientas de aprendizaje automático. Se hizo uso de la estructura de árboles de decisión (DT), cuya evolución se aplicó a la técnica de bosques aleatorios/Random Forest (RF) y Adaptive Boosting (Ada Boost-Adb); la descripción de estas herramientas, métricas y casos en los cuales han sido utilizadas en materia de desarrollo sostenible o calidad del aire puede consultarse en el documento del apéndice C y el análisis de su aplicación puede consultarse en el apéndice A.

Las variables de entrada al modelo fueron las siguientes: 1) contaminantes atmosféricos ($PM_{10,}$ $PM_{2.5}$, SO_2, NO_X, CO); 2) variables meteorológicas (precipitación, temperatura); 3) población que habita en cada UPZ; 4) densidad poblacional según cada UPZ; 5) uso del suelo, cuyas características se detallan en el apéndice A; 6) proximidad de los hogares a las vías primarias (tipo 1:T1) y secundarias (tipo 2:T2).

La información ingresada al modelo corresponde a la distribución espacial de las variables ambientales, representada a través de las salidas gráficas generadas con la aplicación del método IDW para los contaminantes atmosféricos y variables meteorológicas. Para el caso de las variables densidad poblacional, uso del suelo y población, las capas de información espacial están representadas por su distribución en cada unidad de planeación zonal. Adicionalmente, las variables explicativas se representaron en una escala de 0 a 1, en donde los valores cercanos a cero exponían una menor calidad ambiental y/o mayor posibilidad de relación positiva con una afectación a la salud humana. Se ingresó una capa adicional relacionada con la proximidad a vías primarias (T1) y secundarias (T2) y la probabilidad de presentar alguna afectación en la salud respiratoria de la población en la proximidad entre los 100 a 1000 m de los hogares a las vías T1 y T2, de acuerdo con Salam et al. (2008) and Li et al. (2011).

El conjunto de información en formato espacial se agrupó a través de las herramientas de ArcGIS en una sola capa con múltiples bandas de información (ver figura 3-4), generando un consolidado de información vertical de cada variable explicativa y las respuestas categóricas (afirmativas y negativas) del trabajo desarrollado en campo con la identificación de casos asociados a sintomatología y enfermedad del sistema respiratorio.

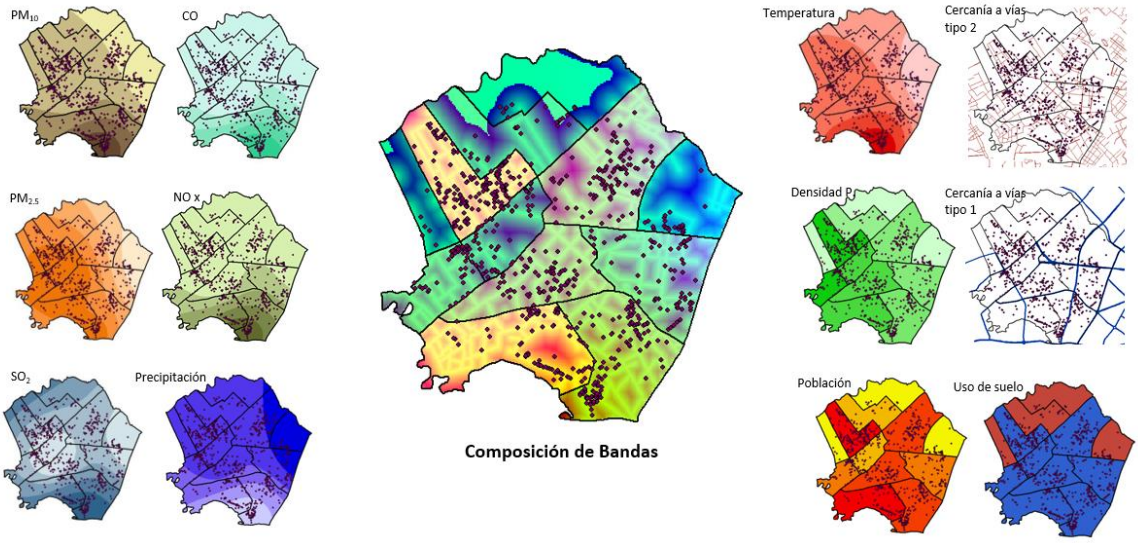


Figura 3-4 Composición de bandas de información.

La calibración del modelo incluyó la iteración de 300-1500 árboles con variaciones de cada 100 árboles, estableciendo la mejor combinación de variables acorde a la exactitud y resultados del índice Kappa. El entrenamiento consistió en hacer uso del 70% de la información de la capa generada y que fue elegida de manera aleatoria para el caso de las variables explicativas; para el caso de la variable respuesta se eligió un 70% de las respuestas afirmativas y negativas a la presencia de enfermedad respiratoria (según el trabajo desarrollado en campo), de tal manera que se presentara consistencia en el entrenamiento y validación del modelo.

El proceso de calibración y entrenamiento consistió en una validación cruzada de partición e iteración de la información de entrenamiento. Esta técnica garantiza la independencia entre los datos de entrada y prueba en el proceso de calibración y ajuste del modelo, permitiendo la generación de un modelo fiable en cuanto a la calidad de la predicción ya que esta técnica aplica de manera iterativa cada modelo sobre todo el conjunto de entrenamiento dividido de manera aleatoria en el conjunto de prueba y de entrenamiento (Kubat, 2017; Rokach y Maimon, 2015).

Para garantizar el uso de datos balanceados en el modelo de predicción se realizó una combinación de las respuestas afirmativas y negativas en dos conjuntos de información: 1) 70% de respuestas que indicaron casos asociados a sintomatología y enfermedad del sistema respiratorio y 70% de las respuestas negativas; 2) el 30% de respuestas afirmativas y el 30% de respuestas negativas. Del primer conjunto se seleccionó de manera aleatoria el 70% de los

datos y del segundo conjunto el 30%; este último fue designado para la validación del modelo.

Como se indicó las herramientas de aprendizaje automático utilizadas fueron Random Forest y Adaboost mediante el uso del software libre R, con métricas de evaluación asociadas a la curva de características de funcionamiento del receptor (ROC, por sus siglas en inglés), exactitud y la medida H, cuya descripción puede consultarse en el apéndice A.

Dado que los datos de entrada al modelo corresponden con información espacial en coordenadas planas, la predicción de la variable respuesta fue ubicada de la misma manera a nivel espacial, según la distribución generada en el pronóstico. La confluencia de las variables explicativas permitió identificar las zonas de mayor interés en el micro territorio. El modelo de pronóstico consistió en un proceso de clasificación de acuerdo con el comportamiento de las variables explicativas y las condiciones de entrada indicadas por la variable respuesta en el proceso de entrenamiento.

3.2 Evaluación del nivel de avance en materia de desarrollo sostenible

El desarrollo sostenible es un objetivo deseable para todos los territorios en el ámbito global. Para la evaluación del nivel de sostenibilidad en el desempeño del micro territorio se llevó a cabo el procedimiento que se esquematiza en la figura 3-5, y se describe a continuación.

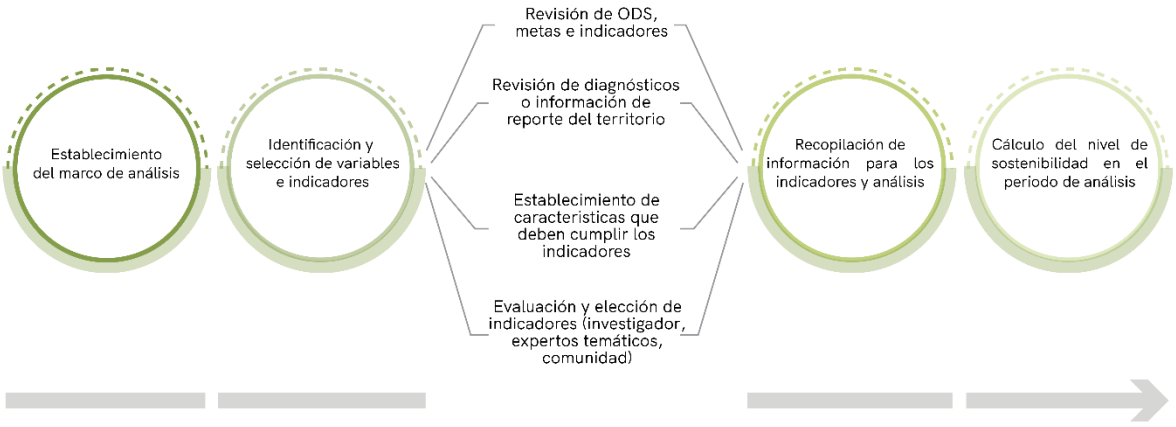


Figura 3-5 Procedimiento general para el cálculo del nivel de sostenibilidad. Adaptado de (Quiroga Martínez et al., 2009)

3.2.1 Establecimiento del marco análisis

Dentro de los marcos de trabajo para analizar el desempeño de un territorio, en términos de sostenibilidad, se destacan el ambiental e indicadores ambientales; el basado en el capital natural e indicadores monetizados y el enfocado en el desarrollo sostenible y sus indicadores (Quiroga Martínez et al., 2009). Considerando la naturaleza del presente trabajo y los fines del desarrollo sostenible, se partió de su conceptualización, el reconocimiento de un contexto internacional y de las políticas fijadas a nivel nacional con relación a los ODS; esto permitió elegir el marco de trabajo que fue el definido por la Comisión de Desarrollo Sostenible de las Naciones Unidas y su marco de ordenador, el cual permite comprender la organización

lógica de los indicadores a partir de unos elementos (Quiroga Martínez et al., 2009), para este caso el basado en temas y subtemas (Organización de las Naciones Unidas, 2018; UN Commission on Sustainable Development, 2001). Los temas o dimensiones rectoras son las denominadas: dimensión ambiental, dimensión social, dimensión económica y dimensión institucional.

Tanto el marco de trabajo y el ordenador presentan unos indicadores ya planteados, que son utilizados en el ámbito internacional y han sido analizados y evaluados en cuanto a su oportunidad por expertos del orden internacional. En este orden de ideas, la inclusión de dichos indicadores para la evaluación del avance en el desarrollo sostenible estaría validada por un marco existente y aplicado en diferentes territorios en el ámbito internacional. Teniendo en cuenta las necesidades y características propias del territorio, se adicionaron en esta investigación algunos elementos de elección e inclusión de posibles nuevos indicadores tal y como se describe en la sección 3.2.2 y que se detalla en el apéndice B.

3.2.2 Identificación y selección de variables e indicadores

El análisis del comportamiento de la sostenibilidad a nivel territorial requiere la elección de variables e indicadores que mejor califiquen el comportamiento del territorio. Del mismo modo, la elección de dichas variables debe considerar el marco político en el ámbito global, así como en el ámbito territorial. Una vez fijado el marco de trabajo y su marco ordenador se procedió a la identificación de variables e indicadores aplicables a una zona urbana, para ello se hizo revisión de los 17 ODS, sus metas e indicadores, así como las metas fijadas a nivel de país y los indicadores e información publicados en observatorios o reportes de entidades públicas.

Se hizo revisión de las quejas y peticiones generadas por la comunidad para el territorio elegido en el periodo de análisis, identificando, a través de un análisis de frecuencias, las principales temáticas objeto de petición, lo cual refleja los aspectos centrales que requieren de monitoreo y medición y cuya incidencia o prevalencia puede revelar una necesidad de mejora. Adicionalmente, se efectuó un análisis a través de minería de texto que permitió identificar la información que se divulga a la comunidad respecto a los ODS, considerando la información digital publicada en las páginas web de los diferentes medios de comunicación nacionales y en el ámbito de la ciudad capital en el periodo 2009-2018. Como parte del preprocesamiento de la información digital se eliminaron palabras vacías, espacios excesivos, números y textos propios del medio de comunicación generador de la información y que no aportaran en el análisis de textos, tal y como se describe en el apéndice D en este documento. Los resultados en el análisis de frecuencia de quejas y peticiones, como los de la minería de textos, fueron elementos orientadores en la identificación de los indicadores a analizar.

De manera paralela se realizó una búsqueda de información asociada a la identificación de indicadores, estableciendo las características de evaluación para cada uno, las cuales se definieron a partir de un conjunto de 35 características analizadas en diferentes estudios. El conjunto de las características que se utilizó para la elección de los indicadores correspondió con: la accesibilidad de información; la solidez analítica; la universalidad, en términos de

escalas espaciales y temporales; la relevancia política y la utilidad para los usuarios; el enfoque multidimensional; la posibilidad de ser un indicador medible; un indicador con características de inequívoco y sistemático.

Se realizó una identificación previa de un primer conjunto de indicadores, que se calificaron a partir de las características elegidas. Posteriormente se realizó consulta a expertos técnicos acerca de la relevancia de involucrar los indicadores para evaluar el nivel de sostenibilidad de la zona urbana, con las características del territorio analizado. Para ello se diseñó un formato de recopilación de información en línea y que se remitió vía web a los expertos temáticos seleccionados de un conjunto de profesionales que laboran en las áreas de planeación, sostenibilidad urbana, salud, ambiente y que son formados profesionalmente en dichos ámbitos. Adicionalmente, teniendo en cuenta que la comunidad de la zona es experta conocedora de su territorio, se le hizo consulta para identificar la importancia de los indicadores, así como aquellos que no se hayan involucrado en el análisis previo. Para este último caso se realizaron, en el año 2018, dos talleres en los cuales se presentó la temática de indicadores de sostenibilidad y se estableció el conjunto de indicadores a evaluar, con base en un formulario que fue diligenciado por los asistentes.

La información recopilada en el desarrollo de los talleres fue procesada y analizada de manera descriptiva identificando el comportamiento de las respuestas generadas por cada actor respecto a los indicadores propuestos en los formatos de trabajo. El concepto de los expertos temáticos y de la comunidad generó la información necesaria acerca de la relevancia de cada indicador evaluado.

3.2.3 Recopilación de información y análisis de comportamiento de los indicadores y sus posibles interacciones

Una vez identificados los indicadores se procedió a recopilar la información para cada caso en los diferentes años del periodo de análisis, a través de documentos de diagnóstico local, información existente en observatorios y las respuestas generadas por entidades del sector público, según las 27 solicitudes escritas realizadas a dichas instituciones. La información recopilada fue organizada digitalmente y se aplicó estadística descriptiva para la determinación de valores máximos, mínimos, desviación estándar, tendencia en el comportamiento de la información. Posteriormente, se realizó una correlación canónica, estableciendo el comportamiento de los indicadores en el periodo 2009-2017 y las posibles interacciones.

El análisis de correlación canónica se orienta a la comparación de dos grupos de variables en un análisis multivariante y permite determinar posibles relaciones entre variables, así como posibles dependencias. Al crear combinaciones lineales entre las variables dependientes e independientes se maximizan las correlaciones entre los grupos de variables (Härdle y Simar, 2014), reflejando la varianza compartida en las combinaciones lineales y no la extraída de las variables. Se trata de una comparación pareada para identificar el comportamiento y relevancia de una variable frente a la otra; en este caso de un indicador frente a otro.

La correlación canónica permitió comparar los indicadores en relación con las diferentes dimensiones; se identificaron posibles relaciones en las interacciones de los indicadores

pertenecientes a las dimensiones social y económica, para la interacción denominada equitativa; las dimensiones ambiental y social, en la interacción denominada habitable y la relación de las dimensiones ambiental y económica para la interacción denominada viable, tal como se presenta en el apéndice B del presente documento.

3.2.4 Cálculo del nivel de sostenibilidad

Para el cálculo del nivel de sostenibilidad se aplicaron las ecuaciones que se resumen en la tabla 3-1.

Tabla 3-1 Ecuaciones para el cálculo del nivel de desarrollo sostenible (Molina-Gómez et al, 2020)

	Ecuación	Variables	Fuente
1	$SDI = (\frac{1}{4}) \sum_{1}^{4} DI$	SDI = Índice de desarrollo sostenible DI = Índices para cada dimensión	(Rajaonson and Tanguay, 2017; Torres-Delgado and López Palomeque, 2018)
2	$I = \sum_{i=x}^{n} (w_i * y_i)$	I = Indicador w_i = Peso relativo de cada indicador y_i = Valor normalizado de cada indicador	(Rajaonson and Tanguay, 2017)
3	$y_t^i = \frac{x_t^i - \min(x_t^i)}{\max(x_t^i) - \min(x_t^i)} \in (0,1)$	y^i_t = Valor normalizado del indicador x^i_t = Dato del indicador para el periodo t min. (x^i_t) = valor mínimo del indicador máx. (x^i_t) = valor máximo del indicador	(Cui et al., 2019; Rajaonson and Tanguay, 2017)
4	$DI = \sum_{1}^{n} (I)$	DI = Índice según dimensión n=Número de indicadores de la dimensión I = Indicador	(Torres-Delgado and López Palomeque, 2018)

Teniendo en cuenta que el marco de análisis de las naciones unidas fija 4 dimensiones (ambiental, social, económica e institucional), se determinó el nivel de importancia w_i de cada indicador dentro de la dimensión a la cual pertenece, mediante la aplicación del proceso analítico jerárquico (AHP, por sus siglas en inglés) y la escala definida por (Saaty, 1987). En este proceso se incluyó la valoración realizada por expertos técnicos y por la comunidad respecto a la importancia de los indicadores. Posteriormente se calculó el nivel de desarrollo sostenible (SDI) para cada año en el periodo de análisis 2009-2017.

Se identificaron valores esperados o valores meta para cada uno de los indicadores a partir de la comparación con el valor límite de cada indicador establecido en documentos de política, regulaciones y metas en el contexto del territorio (país, ciudad, localidad). Cada indicador agregado en su dimensión permitió establecer el nivel de avance para cada una de las dimensiones y en conjunto el nivel de desarrollo sostenible para cada año de análisis en las categorías alto, medio y bajo en una escala de valoración de 0 a 1, siendo 1 el nivel superior de la valoración en la categoría alto y 0 el valor más bajo en la categoría bajo.

3.3 Identificación de las herramientas de aprendizaje automático

Las herramientas de aprendizaje automático permiten identificar patrones de comportamiento de los datos o procesar series de tiempo, así como establecer el comportamiento futuro de las variables o problemas específicos (Molina-Gómez et al. 2021a). Para efectos del pronóstico se realizó una revisión de su aplicación en los ámbitos de interés de este estudio: calidad del aire y desarrollo sostenible (ver apéndice C). Se

identificaron las diferentes herramientas de aprendizaje automático aplicadas, los requisitos de su implementación, la información de entrada en cada escenario, así como las métricas de evaluación de desempeño de las herramientas en los diferentes modelos de clasificación y/o regresión.

Este trabajo permitió identificar la aplicabilidad de las herramientas para los fines de este estudio, así como los requisitos para el desarrollo del modelo de predicción. El detalle del procedimiento aplicado y los resultados se presentan en el apéndice C. Los resultados obtenidos en esta etapa contribuyeron con el alcance del objetivo específico correspondiente a identificar y seleccionar las herramientas de aprendizaje automático aplicables; además los resultados fueron un insumo para el desarrollo de las actividades consistentes en la aplicación de dichas herramientas y en la evaluación de su desempeño.

3.4 Aplicación de herramientas de aprendizaje automático

Como se ha indicado, el desarrollo sostenible es alcanzable a partir de la sinergia entre las acciones que se implementen dentro de cada una de las dimensiones (ambiental, social, económica e institucional); su medición parte de variables e indicadores que permiten establecer el comportamiento de los recursos y componentes en un periodo determinado. Sin embargo, se requiere de una mirada no sólo al comportamiento temporal sino también en el nivel espacial del desarrollo sostenible; por lo cual, no sólo se trata de la identificación del comportamiento en periodos futuros sino también dentro del territorio, cuyo conocimiento reforzará las decisiones a tomar en cada periodo de implementación de medidas y acciones. Estos dos ámbitos de análisis pueden combinarse para establecer el avance en la sostenibilidad en el desarrollo de un territorio.

Para la elección de las herramientas de aprendizaje automático se consideraron los resultados obtenidos en desarrollo de la etapa 3 descrita en el epígrafe 3.3 del presente documento y que se presentan en detalle en el apéndice C. A continuación, se establece el procedimiento seguido en la aplicación de dichas herramientas en la escala temporal (en el periodo 2009-2017) y en la escala espacial (año 2016) para la clasificación de los niveles de sostenibilidad, de acuerdo con el procedimiento esquematizado en la figura 3-1 del presente documento.

3.4.1 Análisis temporal

La aplicación de herramientas de aprendizaje automático requiere que la entrada de información sea suficiente para el entrenamiento y aprendizaje, por lo que dadas las experiencias documentadas en el apéndice C y el análisis confirmado con la información anual para el periodo 2009-2017, se encontró que la alimentación de un modelo de estas características con información anual no es suficiente, pese a que es la medida temporal más utilizada para el reporte de información acerca del comportamiento de los temas y subtemas de las dimensiones de la sostenibilidad. En este orden de ideas, para establecer el pronóstico de los niveles de sostenibilidad se pasó de un volumen de información anual a información mensual. Por lo cual, con las variables de entrada con información disponible en la escala mensual, se calculó nuevamente el nivel de sostenibilidad y se determinaron las correspondientes categorías mediante el uso de las ecuaciones descritas en la tabla 3-1.

Se realizó partición del conjunto de datos en 70% para entrenamiento y 30% para validación y las tres categorías de clasificación según el nivel de sostenibilidad: alto (0.67–1.0), medio (0.34–0.66) y bajo (0.0–0.33) (Molina-Gómez et al., 2020). Se aplicaron herramientas de clasificación para pronosticar los niveles de desarrollo sostenible a partir del uso de los indicadores. Las herramientas aplicadas y comparadas fueron: árboles de decisión (C.5.0Tree), redes neuronales artificiales (algoritmo de perceptrón), y la máquina de vector soporte (SVMradial). La calibración de los modelos obedeció a la iteración automática de parámetros específicos para la operación de cada herramienta de aprendizaje automático, eligiendo las características de los modelos de mejor desempeño a partir de las métricas exactitud e índice Kappa. Se realizó una validación cruzada de tal manera que en la calibración y entrenamiento se estableciera el modelo con el mejor desempeño.

Se hizo uso del software de acceso libre R junto con la librería *caret*, utilizando los siguientes paquetes: el método C5.0Tree (Kuhn et al., 2013) para árboles de decisión, el método y paquete nnet (Ripley y Venables, 2016) para la aplicación de redes neuronales artificiales y la función del paquete e1071 (Meyer et al., 2019) para el modelo en que se aplicó SVM. (Molina-Gómez et al., 2020)

3.4.1.1 Desempeño de los modelos

Las métricas de evaluación de desempeño de los modelos según las herramientas de aprendizaje automático aplicadas fueron: exactitud balanceada, precisión, exhaustividad/sensibilidad y tasa de verdaderos negativos a partir de la información generada por la matriz de confusión. Las variables utilizadas en el cálculo de las métricas se muestran en la tabla 3-2 y que se construyó a partir de las ecuaciones de evaluación de desempeño de modelos de aprendizaje automático descritas por (Dinov, 2018; Kubat, 2017; Rokach y Maimon, 2015).

Tabla 3-2 Métricas de desempeño utilizadas en la evaluación del modelo

Métrica de desempeño	**Elementos de la métrica**
Exactitud balanceada	$\frac{sensibilidad + especificidad}{2}$
Precisión	$\frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos}$
Exhaustividad/sensibilidad (recall)	$\frac{verdadero\ positivos}{verdaderos\ positivos + falsos\ negativos}$
Especificidad	$\frac{verdaderos\ negativos}{verdaderos\ negativos + falsos\ positivos}$

Se destaca que, dadas las condiciones de la zona de estudio, el comportamiento de los datos en cada una de las dimensiones y que algunas categorías del nivel de sostenibilidad son más frecuentes que otras, se aplicó la métrica de exactitud balanceada, ya que los clasificadores con muestras desbalanceadas pueden generar altos niveles de desempeño en las clases más representativas (Gibert et al., 2018).

3.4.2 Análisis espacial

Con el fin de mantener la coherencia metodológica se hizo uso del mismo marco de trabajo utilizado en el análisis temporal, correspondiente al establecido por la Comisión de

Desarrollo Sostenible de las Naciones Unidas, basado en temas y subtemas. No obstante, para el pronóstico en el ámbito espacial se trabajó únicamente con las tres dimensiones: ambiental, social y económica; no se incluyó la dimensión institucional dado que las intervenciones desde la institucionalidad en la zona de estudio obedecen a comportamientos que no varían en el territorio dentro de un mismo periodo de tiempo, siendo de aplicación homogénea para el periodo analizado (2016). No obstante, es necesario revisar para cada caso específico el comportamiento espacial de los indicadores en la dimensión institucional, pues eventualidades relacionadas con la variabilidad climática, posibles riesgos tecnológicos o biológicos, entre otros, pueden hacer que desde la institucionalidad se ejecuten acciones diferenciadas a nivel espacial.

Se aplicó una metodología de evaluación para el año 2016, con el fin de identificar el nivel de desarrollo sostenible en un año en el espacio geográfico analizado, así como establecer la influencia de la calidad del aire en dicho avance. Dada la definición de zonas de interés en términos de calidad del aire y salud realizada previamente, se buscó identificar además el nivel o niveles de sostenibilidad que confluyen en dichas zonas. El modelo de análisis para cada año en el periodo de estudio permite pronosticar el avance global del territorio en general, pero con el análisis espacial es posible profundizar en el comportamiento de los indicadores y dimensiones en el territorio para cada una de las UPZ.

Es del caso mencionar que la información de entrada al modelo en la escala temporal no se encuentra para todos los casos en la misma unidad de medida temporal ni espacial, es así que fue necesario identificar los indicadores aplicables al contexto espacial. Al respecto, se consideraron tres elementos para la definición de los indicadores: 1) el marco de análisis y ordenador establecido, 2) los indicadores que presentaron mayor influencia en los modelos desarrollados previamente en este estudio (véase procedimiento en epígrafes 3.1 y 3.2 y los apéndices A y B), y 3) indicadores relacionados con los establecidos en el análisis anual y que brindaran información en el ámbito espacial.

Para el pronóstico del nivel de desarrollo sostenible en el plano espacial se aplicó el procedimiento que se esquematiza en la figura 3-6 y que se describe a continuación:

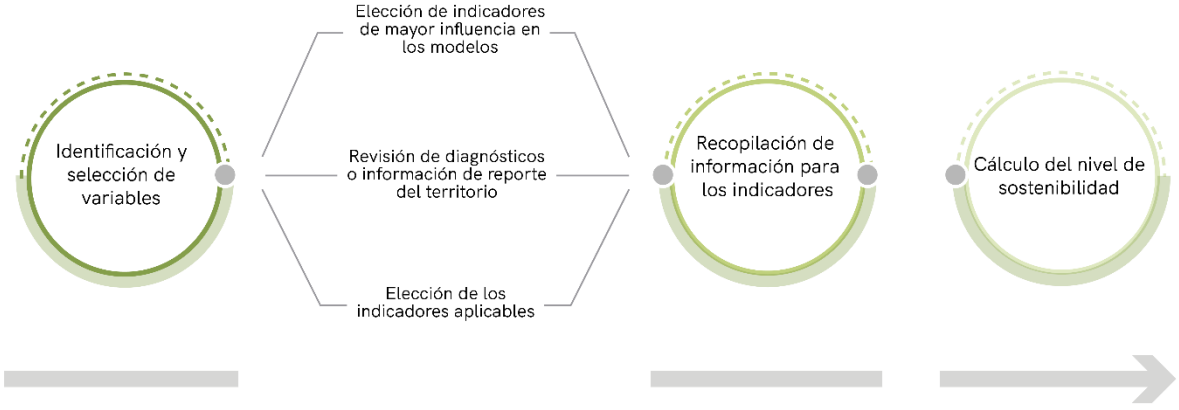


Figura 3-6 Procedimiento general para el cálculo del nivel de sostenibilidad a nivel espacial

3.4.2.1 Selección de indicadores y recopilación de información.

Se hizo una revisión de los indicadores utilizados en la evaluación para cada año del periodo de análisis, identificando aquellos con información disponible o que permitieran generar una descripción de la zona urbana a nivel espacial. Para este caso la recopilación de información de indicadores de la dimensión ambiental, en materia de espacios verdes, superficies impermeables, temperatura superficial terrestre y expansión urbana, se fundamentó en los resultados de los índices espectrales generados a partir del uso de imágenes satelitales Landsat 7 Enhanced Thematic Mapper ETM+ (2000-2013) y Landsat 8 Operational Land Imager OLI (2013-2020) (ver apéndice E). El cálculo del indicador de expansión urbana se realizó a partir de la comparación entre la información existente del año más cercano (2015) al año de análisis y el año 2000, comparando los índices diferenciales de vegetación normalizada (NDVI) del año 2000 y 2015. Es preciso mencionar que no se contó con información para el año 2016 a partir de las imágenes satelitales, dada la alta presencia de nubosidad que no permitió registrar la información requerida para el cálculo de los índices espectrales.

Por otra parte, especial atención se prestó a los indicadores de mayor influencia tanto en la definición de zonas de mayor interés en términos de calidad del aire y salud (ver apéndice A) como en el pronóstico del desarrollo sostenible según el modelo con mejores métricas de desempeño (ver apéndice B). Otros indicadores de la dimensión ambiental fueron establecidos a partir del método de IDW, siguiendo el procedimiento establecido en la sección 3.1.2 del presente documento. La información espacial de estos indicadores se recopiló en el formato *shape* con el fin de realizar la lectura de la información a través del software ArcGis y el pronóstico mediante el uso del lenguaje de programación de acceso libre R.

Los indicadores relacionados con las dimensiones social y económica se fundamentaron en información publicada por entidades gubernamentales, en especial el departamento administrativo nacional de estadística (DANE), en lo relacionado con la medición de pobreza multidimensional (IPM) de fuente censal (DANE, 2020). Para cada indicador se generó una capa con la información espacial relacionada; en el caso de la información del IPM se hizo uso de los factores de ponderación establecidos por el DANE (DANE, 2020) para la determinación de cada indicador de manera diferencial.

La elección de los indicadores, la búsqueda y la recopilación de la información se estableció para contar con capas de información descriptivas del comportamiento espacial de cada uno en el territorio analizado, obteniendo un total de 26 capas asociadas al comportamiento espacial de cada indicador.

En esta etapa se realizó un proceso de análisis de correlación de los indicadores para identificar posibles relaciones entre ellos, como soporte en el análisis de su comportamiento dentro de la dimensión a la que pertenecen y entre dimensiones en el ámbito espacial. Para ello se aplicó el método de correlación de Pearson. Este análisis se constituye en un insumo para comprender el comportamiento de las variables en la identificación de la proporción de aporte en el desempeño del territorio.

3.4.2.2 Cálculo del nivel de sostenibilidad y pronóstico de su comportamiento a nivel espacial

Se siguió el mismo procedimiento establecido en el epígrafe 3.2.4 y la aplicación de las ecuaciones 2, 3 y 4 (ver tabla 3-1). Se aplicó la ecuación 2, en donde se estableció para cada indicador un peso relativo (wi) en relación con el conjunto de indicadores pertenecientes a su propia dimensión; se aplicó la ecuación 3 con el método de normalización de máximos y mínimos, verificando además la naturaleza del indicador y, la ecuación 4 que permitió la obtención del subíndice para cada dimensión. El nivel de sostenibilidad se calculó como el promedio de los subíndices de las dimensiones ambiental, social y económica como se presenta en la ecuación 5 de la tabla 3-3.

Tabla 3-3 Ecuaciones para el cálculo del nivel de desarrollo sostenible con variables espaciales

Ecuación		Variables	Fuente
5	$SDI = (\frac{1}{3}) \sum_{1}^{3} DI$	SDI = Índice de desarrollo sostenible DI = Índices para cada dimensión	Adaptado de (Rajaonson and Tanguay, 2017; Torres-Delgado and López Palomeque, 2018)

La información ingresada al modelo corresponde a la distribución espacial de las variables. El conjunto de información en formato espacial se agrupó a través de las herramientas de ArcGIS en una sola capa, generando un consolidado de información vertical de cada variable explicativa.

Las herramientas de aprendizaje automático utilizadas y comparadas fueron Random Forest (RF), redes neuronales artificiales (ANN) y la máquina de vector soporte (SVM). Se hizo uso del software libre R para la modelación.

Para el entrenamiento de cada modelo, se utilizó 70% de la información de la capa generada y que fue elegida de manera aleatoria para el caso de las 26 variables explicativas. Respecto a la variable respuesta se eligió un 70% de las respuestas asociadas a los niveles alto, medio y/o bajo de sostenibilidad, de tal manera que se presentara consistencia en el entrenamiento y validación del modelo.

La calibración del modelo en el que se utilizó Random Forest incluyó la iteración de 0-1000 árboles con la combinación de 14 predictores en el proceso de decisión, estableciendo la mejor combinación de variables acorde a la exactitud y resultados del índice Kappa (Exactitud= 0.9470; índice Kappa = 0.8643; número de árboles = 500). Por su parte el modelo de redes neuronales utilizó el algoritmo de perceptrón monocapa para el modelo de red neuronal de alimentación hacia adelante; éste se calibró considerando la mejor combinación de la exactitud e índice Kappa (número de neuronas: 5; factor de decaimiento:0.0001; exactitud:0.95; índice Kappa:0.87). Para el caso del modelo de máquina de soporte vectorial se utilizó la función radial y los parámetros de calibración sigma y c (sigma:0.06465; c:1; exactitud: 0.9468; índice Kappa:0.8622). En todos los casos se buscó la mejor combinación de los parámetros de ajuste de los modelos aplicando una validación cruzada, lo que permitió generar modelos con la mejor combinación del índice Kappa y el valor de exactitud.

El modelo generado permitió identificar las zonas con niveles altos, medios y bajos de sostenibilidad en el ámbito espacial del micro territorio.

3.4.2.3 Desempeño de los modelos

El desempeño de los modelos se evaluó a partir de las métricas precisión, exhaustividad, especificidad y exactitud balanceada (ver tabla 3-2), las cuales se determinaron a partir de la matriz de confusión en el proceso de clasificación realizado por cada modelo. Como métrica de desempeño adicional, se determinó la media armónica de los resultados de precisión y sensibilidad en cada modelo.

3.5 Determinación de la influencia de la calidad del aire en el desarrollo urbano sostenible

Una vez aplicadas las métricas de evaluación se determinó en primer lugar el modelo de clasificación con el mejor desempeño y en segundo lugar el grado de importancia de las variables explicativas en la clasificación realizada por cada modelo. Mediante el uso de la función "importance" en cada modelo de clasificación, de acuerdo con el paquete de programación empleado, fue posible determinar la proporción de aporte de las variables en el proceso de clasificación supervisada. Se empleó el índice de Gini, que es una medida de desorden, con el fin de identificar la variabilidad que aportan los predictores a la variable respuesta; se identificaron las características que separan mejor la incertidumbre de la información sobre la característica objetivo, lo cual permitió establecer el grado de influencia de las variables sobre el nivel de desarrollo sostenible pronosticado en el ámbito espacial y temporal (Singh et al., 2010).

3.6 Estructuración de la metodología para la determinación de la influencia de la calidad del aire en el desarrollo sostenible

Finalmente, a partir de los resultados generados en cada etapa del proceso metodológico (ver figura 3-1), definido en función de los objetivos planteados para este estudio, se estructuró la metodología general para la determinación de la influencia de la calidad del aire en el desarrollo urbano sostenible mediante el uso de herramientas de aprendizaje automático. En el diseño metodológico generado, como producto de este trabajo doctoral, se concretaron diferentes elementos procesales e insumos necesarios para dar aplicabilidad en zonas urbanas.

4 Capítulo 4. Resultados y Discusión

A partir de la aplicación de las etapas y actividades descritas en el capítulo 3 Materiales y Métodos, se generaron resultados que permitieron establecer el nivel de influencia de la calidad del aire en el desarrollo urbano sostenible. El procedimiento seguido se concretó en el alcance del objetivo general de este trabajo de doctorado consistente en establecer la incidencia de la calidad del aire en el desarrollo urbano sostenible, mediante el uso de herramientas de aprendizaje automático, como soporte para la toma de decisiones.

La ejecución de las actividades encaminadas al desarrollo de los cinco objetivos específicos planteados, fueron el eje direccional de este trabajo de doctorado. Los resultados de cada objetivo específico fueron instrumentos de decisión para el desarrollo de las siguientes etapas del trabajo, así como insumos de entrada en los procesos de cálculo y modelación realizados. Al final se concretó el desarrollo de los objetivos específicos en cuatro documentos manuscritos, tres de estos publicados como productos de la investigación y uno sometido a revista indexada para posible publicación; adicionalmente se generaron documentos presentados en eventos científicos a partir de los resultados intermedios del trabajo realizado. Estos productos se presentan en los apéndices A al E.

En la siguiente figura se hace una relación entre los objetivos específicos planteados (recuadros en color verde), el procedimiento general desarrollado (línea central de la figura en color naranja), y los productos de investigación generados (recuadros en color azul):

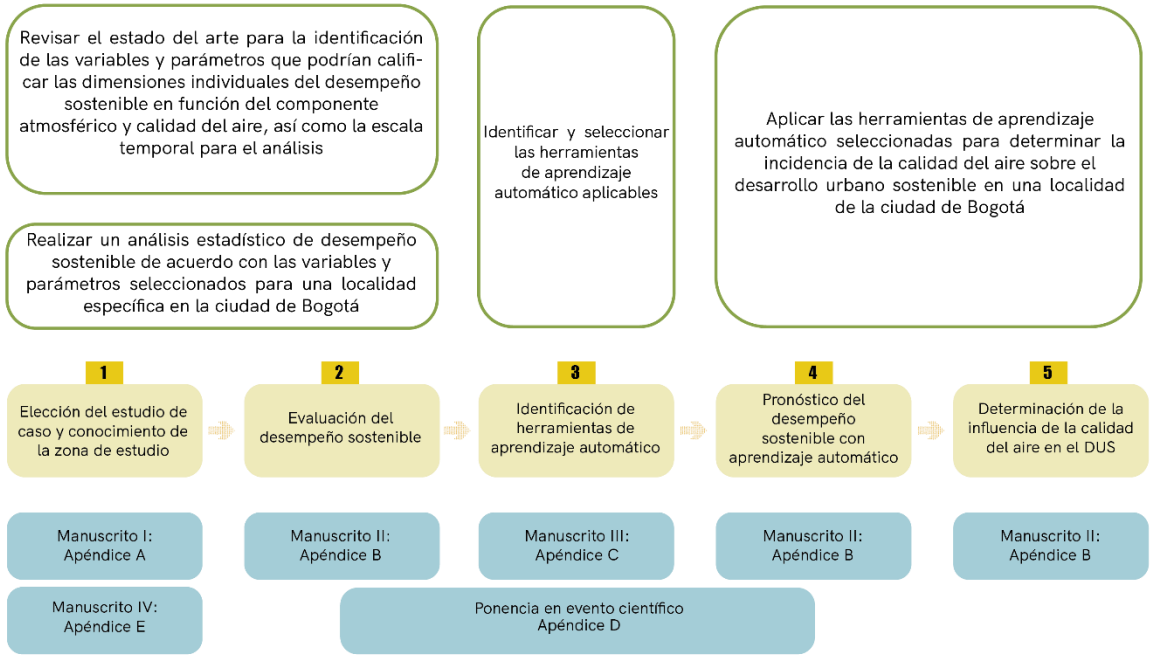


Figura 4-1 Relación de objetivos, etapas de la investigación y productos como resultados de la investigación

A continuación, se presentan los resultados del trabajo doctoral de acuerdo con los objetivos específicos planteados, el esquema metodológico seguido y las publicaciones generadas en cada caso.

4.1 Estado del arte para el análisis del desempeño sostenible

El desarrollo sostenible corresponde a la integración de las dimensiones ambiental, social y económica bajo un marco institucional que soporta dichas dimensiones y sus interacciones. El desempeño sostenible evalúa los resultados de la actuación (Carrillo-Rodríguez y Toca, 2013), es decir los resultados de la aplicación de políticas, estrategias e intervenciones a propósito de los pilares del desarrollo sostenible. A partir de su evaluación en diferentes ámbitos; a nivel empresarial, sectorial, territorial y la comparación de sus resultados se establecen oportunidades de mejora o se reconocen factores claves para la toma de decisiones. Sin embargo, en el contexto urbano la mayoría de los estudios se han enfocado en el análisis de naciones y ciudades capitales, dejando de lado a los micro territorios o espacios geográficos que hacen parte de las ciudades.

Las variables y parámetros para evaluar el desempeño sostenible, así como la definición de la escala temporal tienen como elementos de elección el contexto territorial de análisis. Es por esto que, con base en el marco elegido para la evaluación del desempeño sostenible, se inició con un análisis de la zona establecida como estudio de caso, identificando los indicadores, variables y parámetros más apropiados y la escala temporal de análisis. Ello permitió reconocer el estado del arte en virtud de la calidad del aire, salud y dimensiones de la sostenibilidad en el territorio, así como su desempeño.

4.1.1 Elección del estudio de caso y conocimiento de las características de la zona elegida

La localidad de Kennedy, ubicada en Bogotá, Colombia fue el estudio de caso elegido. Esta es una zona con características de alto interés en lo que respecta a los indicadores de sostenibilidad en las cuatro dimensiones. Es la segunda localidad con mayor número de habitantes en la capital del país, 14.5%; la tercera con mayor extensión en la misma capital (11.62%) y la que ha registrado en los últimos años los mayores niveles de concentración de contaminantes atmosféricos, especialmente PM_{10} y $PM_{2.5}$. Sus índices de pobreza multidimensional (IPM) se encuentran dentro del promedio de ciudad con registros inferiores al 20% y algunos hogares con privaciones que se registran en el rango de 20 a 40% de IPM. En esta localidad se ubica la principal central de abastecimiento de alimentos y víveres del país (CORABASTOS), y se desarrollan diversas actividades económicas. No obstante, los usos de suelo establecidos para cada UPZ, éstos se combinan principalmente con actividades de prestación de servicios, bodegaje de residuos, comercio, y algunas actividades de orden manufacturero.

Para el conocimiento de la zona de estudio en términos de la relación calidad del aire y salud se realizó un trabajo de campo de identificación de casos diagnosticados de enfermedad respiratoria en 2016; el detalle de las actividades efectuadas se desarrolla en el apéndice A. Se encontró que en términos de contaminantes atmosféricos y variables meteorológicas se presentan altos niveles de material particulado en la zona occidental y suroccidental de la

localidad (ver figura 4-2), que coincide con varias UPZ de uso residencial. Este comportamiento es similar para los demás contaminantes atmosféricos con las mayores concentraciones registradas en el año 2016; en la zona sur y occidental de la localidad también se observan los mayores valores de temperatura ambiente.

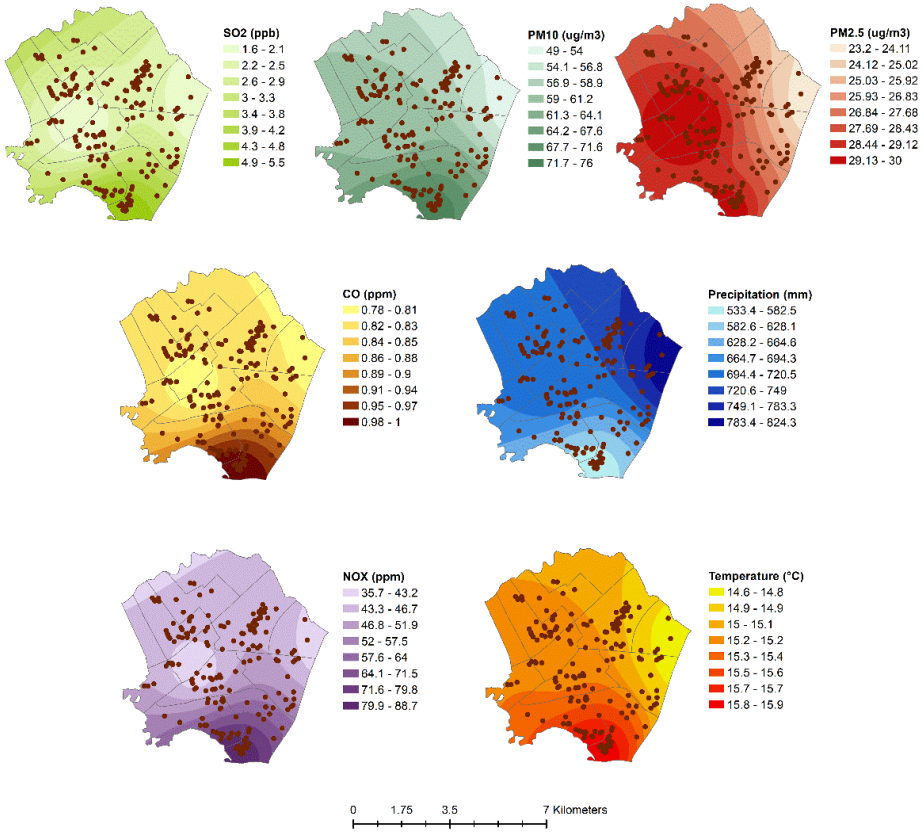


Figura 4-2 Comportamiento de contaminantes, variables meteorológicas y casos de enfermedad respiratoria diagnosticados según trabajo de campo en 2016 (Molina-Gómez et al., 2021a)

El análisis de correlación de las variables material particulado (PM_{10}-$PM_{2.5}$), óxidos de azufre (SO_X), óxidos de nitrógeno (NOx), monóxido de carbono (CO) y temperatura (T) presentó una relación lineal positiva con valores de correlación superiores a 0.8.

Por otro lado, con el fin de concretar en un mejor entendimiento de la zona de estudio se recopiló información de los contaminantes atmosféricos (ver figura 4-3), densidad poblacional, cercanía a vías (primarias T1 y secundarias T2) y uso del suelo. Se aplicaron herramientas de aprendizaje automático en un modelo de predicción de aquellas zonas que podrían presentar mayores posibilidades de enfermedades asociadas a la contaminación atmosférica, principalmente las del sistema respiratorio. Se encontró que las unidades de planeación zonal Patio Bonito y Calandaima son las zonas con una mayor probabilidad de presentar registros asociados a enfermedades del sistema respiratorio. No se descartan casos específicos en UPZ con alta presencia de unidades residenciales como lo son Castilla,

Bavaria, Américas, Kennedy Central y algunos sectores de Timiza y Carvajal (ver figura 4-3).

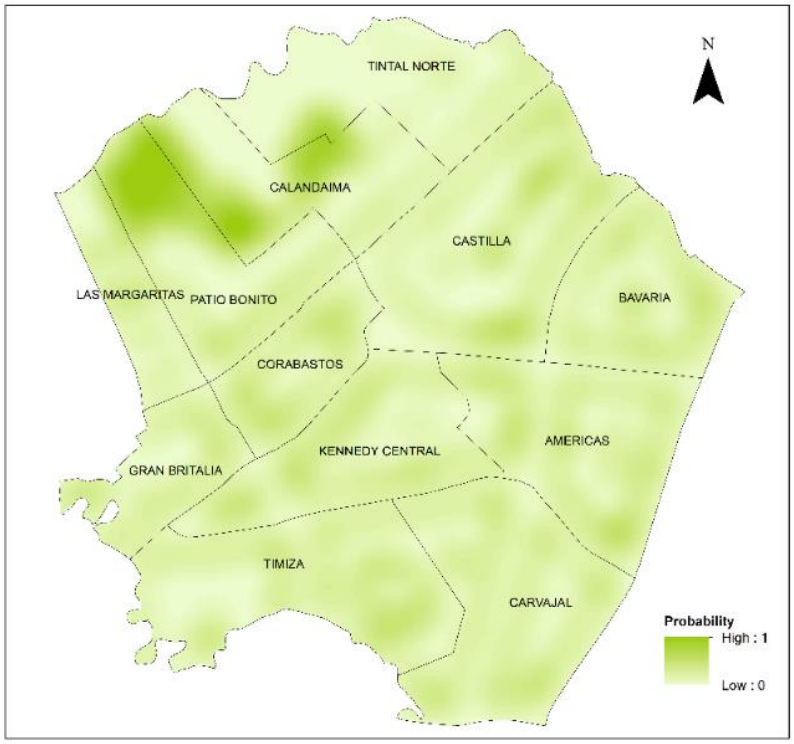


Figura 4-3 Predicción de zonas con posibles casos de enfermedad respiratoria (Molina-Gómez et al., 2021a)

En el análisis se determinaron las variables de mayor influencia en el modelo de bosques aleatorios, siendo éstas la cercanía de los hogares a las vías (primarias -T1 y secundarias -T2), el contaminante $PM_{2.5}$ y la temperatura. La cercanía de las viviendas a las vías soporta los resultados de estudios desarrollados por Li et al. (2011) y Salam et al. (2008) que indican que es un factor asociado a enfermedades del sistema respiratorio; además la emisión de vehículos que operan en las vías y las condiciones de tráfico que se dan en la zona, permiten la exposición de la población a cargas de contaminantes de material particulado fino y ultrafino.

Este modelo presentó una exactitud de predicción de 0.77, con un área bajo la curva (AUC, por sus siglas en inglés) de 0.63 que corresponde con la capacidad de discriminación del modelo, e índice H de 0.1 (ver apéndice A), muy superior al modelo de Adaboost con el cual fue comparado. Estos resultados y el procedimiento seguido para su definición se soportan en el trabajo de investigación publicado bajo la siguiente referencia:

Molina-Gómez, N.I., Calderón-Rivera, D.S., Sierra-Parada, R. et al. Analysis of incidence of air quality on human health: a case study on the relationship between pollutant concentrations and respiratory diseases in Kennedy, Bogotá. Int J Biometeorol 65, 119-132 (2021). https://doi.org/10.1007/s00484-020-01955-4

La localidad de Kennedy es una zona urbana con alta densidad poblacional, en la que se presenta un contraste de edificaciones con el predominio de materiales de construcción en concreto, bloque y ladrillo y alturas superiores a los 8.1 m (cerca del 67% de las edificaciones), otras con alturas superiores a 13.5 m (18.5%) y algunas superiores a los 37 m (Molina-Gómez et al., 2021a). Bajo estas características es posible que se presenten cañones urbanos con la presencia de contaminantes atmosféricos y el fenómeno de la isla de calor urbano.

El fenómeno de las islas de calor urbano se analizó encontrando que en ciertas zonas de la localidad se presentan altos niveles de temperatura superficial terrestre (TST), principalmente en el centro y sur de la localidad; además se encontró que la reducción de zonas verdes y el incremento en la densidad poblacional han sido algunos de los principales factores que han contribuido con dicho fenómeno. Estos resultados y el procedimiento seguido se soportan en el trabajo de investigación sometido para publicación y que se presentan en el apéndice E.

4.1.2 Análisis y evaluación del desempeño sostenible

Antes de realizar la evaluación del desempeño sostenible de la zona de estudio se identificó el conjunto de indicadores que podrían ser aplicables. Como parte del proceso de elección de los indicadores se establecieron dos perspectivas; por un lado, la perspectiva de la comunidad, desde el enfoque de divulgación de información a la comunidad y sus requerimientos (ver figura 4-4) y por el otro, la perspectiva de análisis de un territorio a partir de reportes de comportamiento de indicadores que califican los temas y subtemas en cada dimensión del desarrollo sostenible.

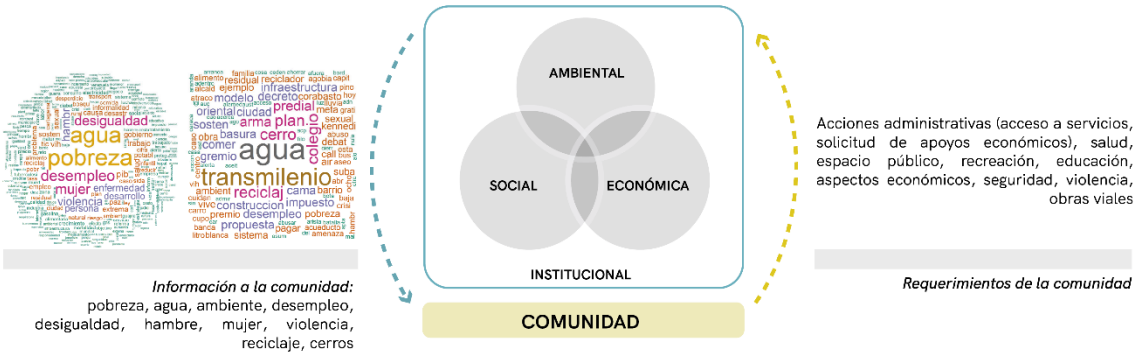


Figura 4-4 Estado del arte respecto a la información que se divulga a la comunidad y requerimientos de la comunidad

4.1.2.1 Análisis de la comunidad

Para el primer caso se eligió a la comunidad como la primera perspectiva porque es el ser humano el eje central de las actividades que se desarrollan en cualquier ámbito y es el fin último del desarrollo sostenible; por lo tanto, los efectos de las acciones que se desarrollen en cualquiera de las dimensiones, inclusive en su articulación, recaerán sobre la comunidad y en el mismo sentido la comunidad es un actor relevante para la implementación de las estrategias establecidas.

Por lo anterior, para establecer el estado del arte, se decidió identificar el grado de información que se emite a la comunidad en materia de desarrollo sostenible, sus objetivos (ODS y del milenio), metas e indicadores a partir de medios de comunicación masiva publicados en medios digitales para el periodo 2009-2018. Adicionalmente, se efectuó un análisis de frecuencia de las peticiones quejas y reclamos de la comunidad en el periodo 2009-2017 con el fin de conocer los requerimientos de la población en dicho periodo.

El análisis de la información publicada en medios se realizó para el contexto nacional y de la ciudad capital en la que se ubica la localidad de Kennedy, mediante la aplicación de herramientas de minería de texto (ver apéndice D). Se encontró que los temas más recurrentes informados a la comunidad tienen que ver con elementos clave de los ODS como pobreza, agua, desigualdad, desempleo (ver figuras 4-4 a y 4-5). En el ámbito nacional, para cada año de análisis, predominaron los términos agua y pobreza, que coinciden con el conjunto de información en el agregado del periodo de 10 años (ver figura 4-5)

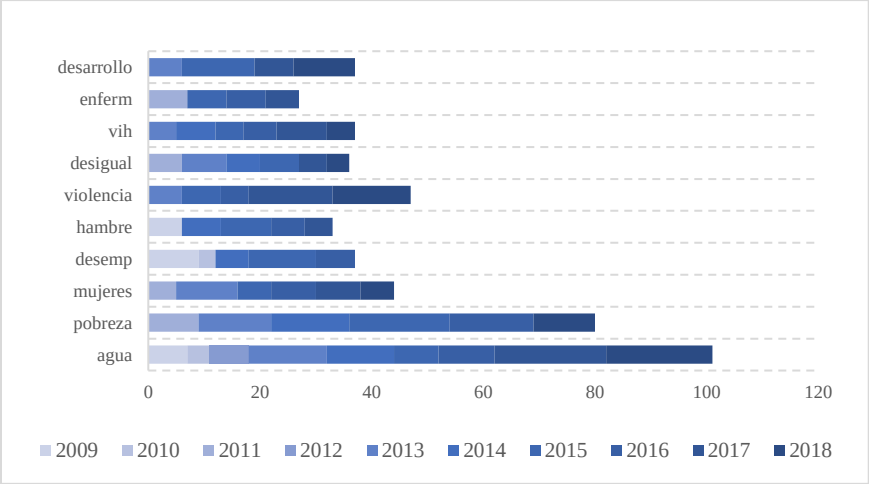


Figura 4-5 Términos más frecuentes en el periodo 2009-2018

Para el término “pobreza”, la información publicada en medios se relaciona con la reducción, salir de la pobreza extrema (17-26% de asociación) y erradicar el hambre (23% de asociación); por su parte, el término “desarrollo” se correlaciona con el término “ecologizar” en un 16%. Para el caso del término “mujer”, se encontró una correlación del 17% con el término “igualdad”, siendo este uno de los resultados más altos en el análisis de asociación de términos. Se resalta que, para los términos relacionados con el componente atmosférico y su calidad, los más cercanos fueron los términos “reforestar” y “morir”, cada uno con un 35% de asociación. En el caso del término “agua”, la correlación es del 51% con el término “potable”, 26% “cobertura”, 17% “escasez” y un 46% con el término “residual”; en este contexto, la información publicada a nivel de ciudad presenta relación con el tratamiento de agua residual (55%).

Se encontró poca cohesión de la información relevante en medios de comunicación con los fines globales del desarrollo sostenible; siendo frecuente la entrega de información puntualizada y de baja articulación con los ámbitos de sostenibilidad.

Se destaca en el análisis de textos la relevancia en el preprocesamiento de la información de entrada, pues una inapropiada limpieza de la información puede ocasionar, más allá de tiempos adicionales en el análisis de los resultados, conclusiones erróneas. Los resultados de la minería de textos y que soportan las temáticas de relevancia pueden encontrarse en el apéndice D o en la siguiente publicación:

N. I. M. Gómez, C. A. Rodríguez, P. A. López and J. L. D. Arévalo, "Minería de texto y aprendizaje automático para identificar prioridades de desarrollo sostenible," *2019 Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASP)*, Barranquilla, Colombia, 2019, pp. 1-5, doi: 10.1109/CASAP.2019.8916682.

Por otro lado, respecto a los resultados del análisis de frecuencias a las solicitudes y quejas, radicadas por la comunidad de la zona de estudio ante entidades públicas en el periodo 2009-2017, se encontró que toman especial relevancia los temas en la dimensión social y económica desde un marco institucional. Las peticiones se enfocan principalmente en aspectos mínimos de seguridad, educación, salud y gestión de obras públicas; para el caso de las temáticas de orden ambiental se encontró que éstas no son las más suscitadas. Dichos requerimientos se encontraron necesarios en la elección y priorización de los indicadores en el territorio.

4.1.2.2 Identificación, calificación y elección de los indicadores (análisis temporal y análisis espacial)

El conjunto de indicadores elegido para la evaluación del desempeño sostenible de la zona de estudio se presenta en la tabla 4-1; en ella se muestra el número de los indicadores de acuerdo con: la dimensión (ambiental, social, económica e institucional), la intersección (habitable, equitativo, viable y sostenible), y el tema general. Adicionalmente se presenta en la misma tabla los ODS con los cuales se relaciona cada uno de los indicadores según la dimensión y tema al que pertenecen.

Tabla 4-1 Conjunto de indicadores elegidos para el análisis de la zona de estudio. Adaptado de (Molina-Gómez et al., 2020)

Item	***Dimensión ambiental***	***Dimensión social***	***Dimensión económica***	***Dimensión institucional***
Tema	*Aire = 5* *Agua = 4* *Residuos = 2* *Espacios verdes = 3*	*Salud = 20* *Educación = 6* *Demografía = 9* *Seguridad = 4* *Cobertura de servicios públicos = 6* *Transporte = 2*	*Estructura económica = 2* *Pobreza = 3* *Consumo y producción = 4* *Ingresos y gastos = 4* *Empleo = 3*	*Gobierno = 3* *Servicios sociales comunales = 1*
Indicadores en cada dimensión	***Ambiental= 14; Social=47; Económica= 16; Institucional=4***			
ODS relacionados	▪ *Agua limpia y saneamiento* ▪ *Ciudades y comunidades sostenibles* ▪ *Vida de ecosistemas terrestres*	▪ *Fin de la pobreza* ▪ *Cero Hambre* ▪ *Salud y Bienestar* ▪ *Educación de calidad* ▪ *Equidad de género* ▪ *Agua limpia y saneamiento* ▪ *Industria, innovación e infraestructura* ▪ *Ciudades y comunidades sostenibles* ▪ *Paz, justicia e instituciones sólidas*	▪ *Fin de la pobreza* ▪ *Salud y bienestar* ▪ *Energía asequible y no contaminante* ▪ *Trabajo decente y crecimiento económico* ▪ *Ciudades y comunidades sostenibles*	▪ *Fin de la pobreza* ▪ *Asociación para lograr metas*
Indicadores en cada intersección	***Habitable=16; Equitativo = 22; Viable=2; Sostenible= 7***			

Por otra parte, en las tablas 4-2 a 4-5 se presentan cada uno de los indicadores por dimensión, las interacciones a las que pertenecen, así como los valores máximos (máx.), mínimos (mín.), promedio y desviación estándar, según la información recopilada para cada año del periodo de análisis.

En la dimensión ambiental (ver tabla 4-2) están presentes los indicadores relacionados con los siguientes ODS: 6) agua limpia y saneamiento, 11) ciudades y comunidades sostenibles y 15) vida de ecosistemas terrestres. Del mismo modo se ubican indicadores de la interacción denominada habitable, es decir, indicadores que establecen un análisis de información de las interacciones entre las dimensiones ambiental y social, y son señalados con el símbolo (*); también se presentan los indicadores que aportan al reto de la sostenibilidad, los cuales están señalados en la tabla 4-2 con el símbolo (****).

Tabla 4-2 Indicadores en la dimensión ambiental para el análisis y evaluación de la zona urbana

Subtema	**Cod**	**Indicador**		**Análisis de los indicadores**			
				Media	**Desviación estándar**	**Max**	**Min**
Aire	E1	Media anual de PM_{10} en la estación Kennedy ($\mu g/m^3$)	(*)	65.6	8.2	78.7	53.1
	E2	Media anual de $PM_{2.5}$ en la estación Kennedy ($\mu g/m^3$)	(*)	28.0	3.2	35.1	24.3
	E3	Excedencias a la norma de calidad del aire para PM_{10}(%)	(*)	4.4	5.3	14.7	0.4
	E4	Excedencias a la norma de calidad del aire para $PM_{2.5}$ (%)	(*)	3.0	5.4	17.0	-
	E5	Índice de calidad del aire en Kennedy	(*)	85.2	8.9	102.0	74.0
Residuos	E6	Residuos dispuestos en relleno sanitario (Ton)	(*)	331,563.9	14,493.9	349,430.3	302,437.1

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
	E7	Residuos de construcción y demolición con disposición apropiada (RCD) (Ton)	(*)	8,204.5	1,084.4	9,478.4	6,052.8
Agua	E8	Aguas residuales tratadas (%)	(*)	-	0	0	0
	E9	Calidad del agua en el río Fucha (tercer tramo)	(*)	35.0	6.5	48.0	27.0
	E10	Calidad del agua en el río Tunjuelo (cuarto tramo)	(*)	42.6	5.1	53.0	35.0
	E11	Índice de calidad del agua	(*)	99.8	0.2	100.0	99.3
Espacios verdes	E12	Áreas de espacios verdes y recreación (km^2)	(*)	0.7	0.0	0.7	0.7
	E13	Número de árboles por hectárea	(*)	30.5	2.1	33.8	27.7
	E14	Áreas protegidas (km^2)	(****)	3.8	0.0	3.8	3.8
Interacción: habitable (*), sostenible (****). Los ODS relacionados son: 6, 11 y 15							

Para el caso de la dimensión social se estableció un conjunto de 47 indicadores que se encuentran relacionados con los siguientes ODS: 1) fin de la pobreza, 2) cero hambre, 3) salud y bienestar, 4) educación de calidad, 5) equidad de género, 6) agua limpia y saneamiento, 9) industria, innovación e infraestructura, 11) ciudades y comunidades sostenibles y 16) paz, justicia e instituciones sólidas. Del conjunto de indicadores que pertenecen a la dimensión social, 16 hacen parte además de la interacción denominada equitativa, señalados en la tabla con el símbolo (**); tres de la interacción habitable, señalados en la tabla con el símbolo (*) y cuatro en el reto de la integración denominado sostenible, señalados en la tabla con el símbolo (****), (ver tabla 4-3).

Tabla 4-3 Indicadores en la dimensión social para el análisis y evaluación de la zona urbana

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
Salud	S1	Tasa de mortalidad por enfermedad cardiopulmonar, enfermedades del sistema circulatorio y otras enfermedades del corazón		10.4	2.1	13.8	7.6
	S2	Tasa de mortalidad por enfermedades crónicas		2.6	0.2	2.9	2.1
	S3	Tasa de mortalidad por neumonía en adultos mayores de 64 años		79.5	11.7	106.9	66
	S4	Tasa de mortalidad por neumonía en menores de 5 años		7.4	4.2	16	2.3
	S5	Tasa de mortalidad por infecciones respiratorias agudas graves para todas las edades		10.5	1.5	14	8.6
	S6	Tasa de mortalidad infantil por todas las causas		10.1	0.9	11.6	8.9
	S7	Desnutrición aguda en niños menores de 5 años (proporción)		1.5	0.1	1.8	1.2
	S8	Desnutrición crónica en niños menores de 5 años (proporción)		15.8	1.3	18.9	14.6
	S9	Desnutrición global en niños menores de 5 años (proporción)		0.9	1.2	3.5	0
	S10	Tasa de mortalidad en niños menores de 5 años (proporción)		11.4	1.0	12.9	10.1
Demografía	S11	Tasa de fertilidad		36.0	12.5	44.0	4.6
	S12	Tasa cruda de mortalidad		3.2	0.2	3.3	2.8
Salud	S13	Tasa de mortalidad por meningitis bacteriana		0.4	0.2	0.9	0.1
	S14	Tasa de mortalidad en niños menores de 5 años por infección respiratoria aguda		4.9	2.9	11.5	1.2
	S15	Tasa de mortalidad por dengue		0.0	0.0	0.2	0
	S16	Tasa de mortalidad por enfermedad diarreica en niños menores de 5 años		0.6	0.6	1.3	0

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
	S17	Mortalidad infantil por cada 1.000 nacidos vivos		151.8	20.9	181	121
	S18	Tasa de mortalidad perinatal		10.0	0.8	11.3	8.9
	S19	Tasa de mortalidad materna		17.2	5.2	25.4	12.3
	S20	Tasa de Hepatitis B por 100,000 habitantes		3.6	1.2	4.8	0.7
	S21	Tasa de mortalidad en lactantes		29.3	13.3	45.6	11
Educación	S22	Promedio de años de escolaridad	(**)	9.4	0.82	10.7	8.3
	S23	Tasa de asistencia escolar	(**)	87.0	4.1	93.5	81.5
	S24	Tasa bruta de cobertura de educación	(**)	84.0	6.6	90.8	72.1
	S25	Tasa de deserción	(**)	2.6	0.3	2.9	1.9
	S26	Población con educación primaria y secundaria	(**)	189,040.9	43266.7	251,374.1	107,751
	S27	Tasa de analfabetismo	(**)	1.4	0.2	1.6	1.1
Demografía	S28	Esperanza de vida al nacer (años)	(****)	76.9	0.5	77.1	76
	S29	Tasa de natalidad		14.2	1.7	16.1	11
	S30	Tasa de envejecimiento de la población	(****)	10.1	1.3	11.9	8.4
Seguridad	S31	Número de muertes por arma de fuego, homicidios, peleas y/o enfrentamientos	(**)	108.8	36	183	60
	S32	Reportes de violencia y abuso doméstico, familiar e infantil (número)	(**)	1,253	1332.7	3,491	258
	S33	Robo simple y agravado (número)	(**)	3,148.9	1652.7	6,651	1,549
	S34	Robo de vehículos (número)	(**)	531.7	109.24	697	416
Transporte	S35	Pasajeros transportados por el sistema de transporte público masivo (millones)	(****)	49.2	81.9	61.3	36.4
Demografía	S36	Densidad poblacional por km^2	(****)	3.0	0.25	3.4	2.6
	S37	Población ubicada en áreas propensas a inundaciones		1,120,115.7	61253.2	1,208,980	1,030,903
	S38	Zonas propensas a inundación (km^2)	(*)	38.5	0	38.5	38.5
Transporte	S39	Muertes por accidentes de tránsito (número)	(*)	3.78	2.5	8.0	1.0
Salud	S40	Suicidios (número)		29.2	7.9	47	19
Demografía	S41	Asentamientos informales (km^2)	(*)	0.0	0.0	0.0	0.0
Cobertura de servicios	S42	Hogares con acceso al servicio de gas natural (número)	(**)	60,209.9	22027.26	90,390	25,777
	S43	Población con acceso al servicio de gas natural en el hogar	(**)	1,023,903.9	99248.2978	1,186,010	937,698
	S44	Cobertura del sistema de drenaje de aguas pluviales (%)	(**)	98.2	1.7	99.7	94.4
	S45	Proporción de la población que utiliza servicios de agua potable gestionados de forma segura	(**)	100	0.0	100	100
	S46	Población con servicios de alcantarillado (%)	(**)	99.9	0.0	100.0	99.9
	S47	Acceso a los servicios de salud (población)	(**)	3,181,801.8	313,139.19	3,567,041	2,556,316
Interacción: Habitable (*), Equitativo (**), Sostenible (****). Los indicadores no marcados no hacen parte de las interacciones y pertenecen puntualmente a la dimensión social. Los ODS relacionados son 1 a 6, 9, 11 y 16							

Los indicadores que hacen parte de esta dimensión describen las necesidades establecidas por la población a través de sus quejas y peticiones, pero también los aspectos priorizados por expertos técnicos y población conocedora del territorio.

Un análisis de correlación canónica entre los indicadores pertenecientes a la dimensión ambiental y social permite mostrar la cercanía entre indicadores asociados a calidad del aire, eventos de mortalidad en población vulnerable y mortalidad debida a enfermedades del sistema cardiopulmonar. Del mismo modo se relacionan eventos de salud con condiciones físicas en las que se ubican los hogares en el territorio, el acceso a servicios como gas natural y sistema de transporte masivo y educación. El análisis de correlación muestra la interacción de los efectos y condiciones del territorio en términos ambientales y de alcance en la

prestación de servicios y/o atención a la población en la satisfacción de necesidades básicas (ver figura 4-6 y tablas 4-2 y 4-3).

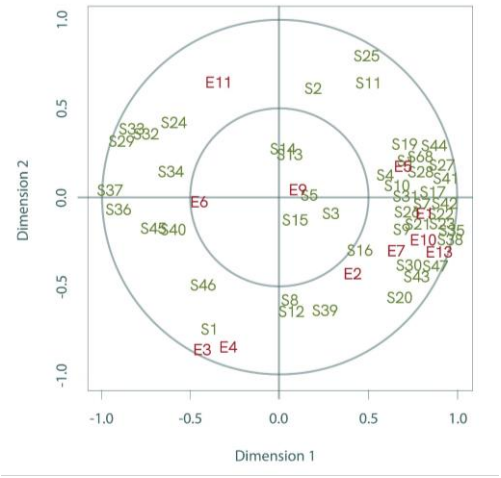


Figura 4-6 Correlación canónica entre los indicadores de la dimensión ambiental y social (Molina-Gómez et al., 2020)

Para el caso de la dimensión económica se eligió un conjunto de 16 indicadores (ver tabla 4-4), de los cuales cinco hacen parte de la interacción denominada equitativa; dos indicadores de la interacción denominada viable, los cuales están señalados en la tabla con el símbolo (***) y, un indicador hace parte de la interacción denominada sostenible. Los indicadores que hacen parte de la dimensión económica están enfocados en los siguientes ODS: 1) fin de la pobreza, 3) salud y bienestar, 8) energía asequible y no contaminante y el ODS 11 ciudades y comunidades sostenibles. La identificación específica de cada indicador respecto del ODS relacionado puede consultarse en el material suplementario del apéndice B.

Tabla 4-4 Indicadores en la dimensión económica para el análisis y evaluación de la zona urbana

Subtema	**Cod**	**Indicador**		**Análisis de los indicadores**			
				Media	**Desviación estándar**	**Max**	**Min**
Ingresos y gastos	EC1	Producto interno bruto per cápita		0.5	0.1	0.7	0.3
Empleo	EC2	Tasa de desempleo	(**)	7.4	0.0	7.5	7.3
	EC3	Población empleada de 12-64 años de edad		511,454.0	43,638.9	571,821.0	432,001.0
	EC4	Población económicamente activa		549,042.2	45,734.4	620,053.0	483,082.0
Ingresos y gastos	EC5	Ingreso familiar per cápita (COP)		807,207.3	171,231.1	981,690.0	567,895.0
Estructura económica	EC6	Afiliación al sistema general de seguridad en salud (%)		88.6	2.0	91	85
	EC7	Acceso a la electricidad (población)	(**)	1,059,005.1	64182.9	1,208,980.0	999,693.0
Pobreza	EC8	Población bajo línea de pobreza	(**)	115,816.9	52518.7	183,966	45,851
	EC9	Necesidades básicas insatisfechas	(**)	6.5	3.2	14.2	4.4
	EC10	Población en alto riesgo debido a la escasez de agua	(**)	114.2	92.6	233.0	0
Ingresos y gastos	EC11	Impuestos sobre bienes muebles e inmuebles (COP)		119,154,674.1	46,178,337.3	182,989,654.0	55,500,820.0
Consumo y producción	EC12	Consumo energético (millones COP)		251,430.4	48000.27	320,879.3	182,177.9
	EC13	Energía consumida (kWh/año)	**(***)**	684,543,112.2	45,425,947.8	740,591,385.0	614,859,738.0
Ingresos y gastos	EC14	Coeficiente de Gini		0.4	0.024	0.5	0.4

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
Producción y consumo (transporte)	EC15	Vías arterias en buenas condiciones (km)	(***)	207.5	105.26	398.4	108.6
Consumo y producción	EC16	Consumo per cápita de agua (residencial) (m^3/hab-d)	**(****)**	0.1	0.0	0.1	0.1
Interacción: Equitativo (**), Viable (***), Sostenible (****). Los indicadores no marcados no hacen parte de las interacciones y pertenecen puntualmente a la dimensión económica. Los ODS relacionados son 1,3,7,8 y 11							

Un análisis de correlación canónica entre las dimensiones social y económica permitió mostrar la relación entre indicadores que determinan el nivel de acceso a servicios públicos, de educación y la población económicamente activa, lo cual establece un relacionamiento entre aspectos como la inclusión social y el crecimiento económico, en subtemas como pobreza, educación y cobertura de servicios públicos (ver tablas 4-3 y 4-4 y figura 4-7).

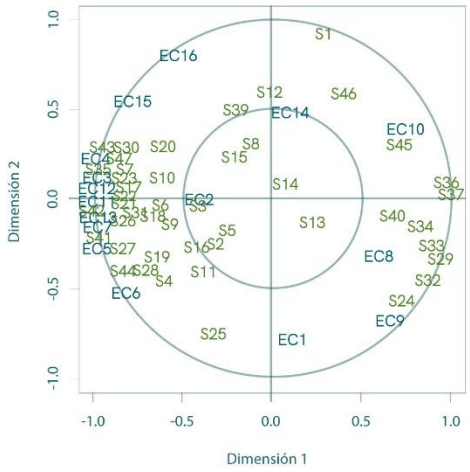


Figura 4-7 Correlación canónica entre los indicadores de la dimensión social y económica (Molina-Gómez et al., 2020)

Adicionalmente, al analizar la dimensión ambiental y económica con ayuda de una matriz de correlación canónica (ver figura 4-8 y tablas 4-2 y 4-4) se observa la cercanía entre indicadores que califican los km de vías en buenas condiciones y la concentración de $PM_{2.5}$; también se relacionan indicadores de crecimiento económico (ingresos per cápita por hogares y consumo energético) con indicadores de presión ambiental asociados a la concentración de PM_{10}, árboles por hectárea y calidad del agua en el tramo de río Tunjuelo, indicadores que describen las condiciones urbanas de la zona de estudio.

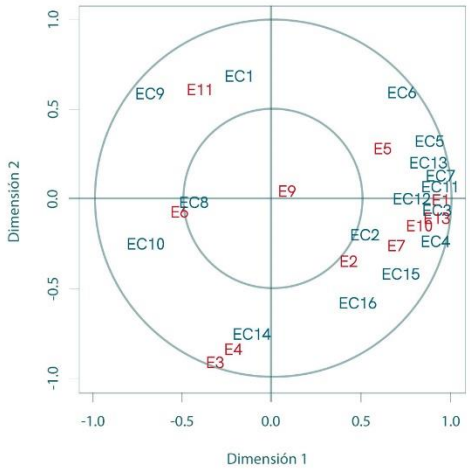


Figura 4-8 Correlación canónica entre los indicadores de la dimensión ambiental y económica (Molina-Gómez et al., 2020)

Con el análisis de correlación canónica entre las dimensiones ambiental, social y económica fue posible establecer la estructura óptima en la cual las variables dependientes e independientes maximizan su relación. Se destaca que se trata de variables que en muchos casos no tienen ninguna relación claramente definida, pero con la correlación canónica fue posible identificar las variables que podrían estar correlacionadas; lo cual da indicios de comportamiento de los indicadores. Este análisis multivariado supera el desarrollo de un análisis de componentes principales dado que en este último unas pocas variables explican la mayor parte de la variabilidad de los datos de entrada.

La dimensión institucional busca evaluar el soporte desde el gobierno en el marco de instrumentos para el impulso y desarrollo de las metas asociadas a los ODS; en este sentido se incluyeron cuatro indicadores, de los cuales uno de ellos hace parte de la interacción denominada equitativa y uno aporta al reto de la integración denominada sostenible (ver tabla 4-5). Los ODS relacionados son: 1) fin de la pobreza y 16) asociación para lograr las metas.

Tabla 4-5 Indicadores en la dimensión institucional para el análisis y evaluación de la zona urbana

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
Gobierno	I1	Políticas o estrategias de desarrollo sostenible (número de políticas)	(****)	5.2	0.8	7.0	4.0
	I2	Gasto público en educación (millones COP)	(**)	268,930	57,000.4	391,475.2	216,624.8
	I3	Tasa bruta de participación		58.6	4.4	62.1	48.4
Servicios sociales y comunales	I4	Eventos culturales (número)		2,228	154.1	2,480	2,008
Interacción: Equitativo (**), sostenible (****). Los indicadores no marcados no hacen parte de las interacciones y pertenecen puntualmente a la dimensión institucional. Los ODS relacionados son 1 y 16							

La información anual de los indicadores permitió determinar el nivel de avance en el desempeño sostenible de la zona de estudio para cada año del periodo elegido, de acuerdo con la aplicación de las ecuaciones de la tabla 3-1 y cuyos resultados se muestran de manera

gráfica en la figura 4-9. De acuerdo con ello, se observa cómo a partir del 2016 existe una tendencia a superar el nivel medio de los rangos establecidos para el nivel de desarrollo sostenible en el territorio.

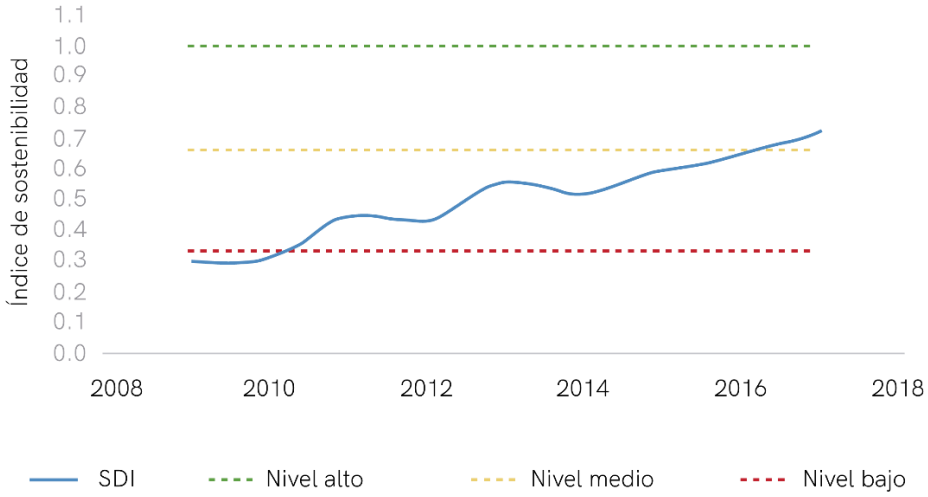


Figura 4-9 Nivel de sostenibilidad para cada año en el periodo y caso de estudio (Molina-Gómez et al., 2020)

En el tiempo un mejor desempeño sostenible estará soportado por una mejora de las condiciones ambientales, sociales y económicas del territorio analizado; para cada dimensión existe una influencia de diversos factores que deben ser reforzados, es decir requieren ajuste, requieren inversión en estrategias que orienten la sostenibilidad de las mejoras que se evidencian en un momento determinado. Por lo tanto, un futuro resultado favorable acorde con los indicadores planteados, para la evaluación del desempeño del territorio, significa que las condiciones se modifican. Sin embargo, se requiere de una evaluación detallada al interior de cada dimensión que complemente el resultado global y no se muestren como un único resultado.

El comportamiento anual de los indicadores refleja información global y generalizada para el territorio. Como se indicó en la sección 3.1.1 de este documento, la localidad presenta características territoriales que varían en virtud de los usos de suelo y las actividades que se desarrollan, es por esto que fue necesario analizar el comportamiento de las dimensiones de la sostenibilidad a nivel espacial.

Considerando que la disponibilidad de información es una de las principales limitantes para evaluar el avance en el desarrollo sostenible de un territorio, se hizo una nueva revisión de los indicadores, seleccionando para cada dimensión aquellos que brindaran información de su comportamiento a nivel espacial. Los factores que influenciaron la elección de los indicadores se basaron en los parámetros de elección analizados en el numeral 3.2.2 del presente documento y las conclusiones del análisis de peticiones de la comunidad y la valoración de expertos. Además, fueron de especial interés las variables de mayor influencia presentadas en el estudio de caracterización de la localidad (ver epígrafe 4.1.1 y el apéndice

A), y las de mayor influencia para el pronóstico del desarrollo sostenible en la escala temporal (ver epígrafe 4.3.1.1 y el apéndice B); no obstante, uno de los factores más decisivos correspondió a la disponibilidad de información. En las Tablas 4-6 a 4-8 se presentan las estadísticas correspondientes a los indicadores seleccionados.

Tabla 4-6 Indicadores en la dimensión ambiental para el análisis espacial

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
Ambiental	Ndvi	Espacios verdes	(*)	0.26	0.2	1	-0.12
	Ndisi	Superficies impermeables	(*)	0.03	0.09	0.4	-0.70
	TST	Temperatura superficial terrestre TST	(*)	27.50	2.22	34.27	16.56
	PM_{10}	Media anual de PM_{10}	(*)	58.88	4.52	76.1	52.57
	$PM_{2.5}$	Media anual de $PM_{2.5}$	(*)	27.43	1.69	30	23.28
Interacción: Habitable (*) Los ODS relacionados son 11 y 15							

El indicador de espacios verdes permite orientar un análisis en lo relacionado con indicadores como áreas de espacios verdes y recreación (E12), número de árboles por hectárea (E13) y áreas protegidas (E14), los cuales hace parte del tema espacios verdes descritos en la tabla 4-2 para los indicadores anuales.

Del mismo modo, indicadores como el denominado superficies impermeables y la temperatura superficial terrestre (TST) dan cuenta de las variaciones en el ámbito espacial que cada año pueden reflejar variaciones limitadas, pero en el territorio pueden presentar diferentes comportamientos (ver apéndice E). El mismo caso se presenta para los contaminantes PM_{10} y $PM_{2.5}$, analizados en términos de la media anual y que califican el tema aire en el marco de la dimensión ambiental y se incluyeron además en el análisis temporal. Estos indicadores se reportaron como relevantes en el ejercicio de caracterización de la zona de estudio en términos de calidad del aire y salud respiratoria (ver apéndice A), lo cual los hizo elegibles en el modelo de análisis espacial.

Para los demás indicadores de la dimensión ambiental se encuentra que existen indicadores que presentan un comportamiento constante en todo el territorio, es el caso de las aguas residuales tratadas. Por otra parte, no se cuenta con información espacial que describa el comportamiento de los demás indicadores por lo cual no se incluyeron en el análisis.

Los indicadores que califican a la dimensión económica (ver tabla 4-7) se limitaron a indicadores propios de los hogares y de cambios observados en el territorio, el indicador trabajo informal responde al subtema empleo, el indicador tasa de dependencia económica al subtema ingresos y gastos y el indicador que evalúa el aseguramiento en salud responde al subtema estructura económica, el cual se relaciona con el indicador EC6 de la tabla 4-4. La expansión urbana se integra como indicador debido a las demandas y presiones económicas que el crecimiento espacial de la localidad puede requerir.

Tabla 4-7 Indicadores en la dimensión económica para el análisis espacial

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
Económica	Trab_infor	Trabajo informal	(**)	0.84	1.76	10	0
	TDE	Tasa de dependencia económica		0.84	1.76	10	0

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
	2000-2015 (CUS)	Expansión urbana		0.08	0.2	1.14	0.66
	S_A_Salu	Sin aseguramiento a salud	(**)	0.84	1.76	10	0
Interacción: Equitativo (**) Los indicadores no marcados no hacen parte de las interacciones y pertenecen puntualmente a la dimensión económica. Los ODS relacionados son 8 y 11							

Para tres de los indicadores económicos la fuente de información corresponde a la capa de generada por el DANE en el cálculo de la incidencia de la pobreza multidimensional de fuente censal (DANE, 2020). En esta dimensión los indicadores denominados: trabajo informal, tasa de dependencia económica y aseguramiento a salud se mueven en el rango intercuartílico 0.82 (0;0.82) correspondiente al comportamiento en cada una de las manzanas y unidades de planificación zonal que componen la localidad en estudio. Para el caso del indicador expansión urbana, su información se presenta en el rango intercuartílico 0.18 (-0.02; 0.15) dada la expansión en cuanto a crecimiento de zonas grises sobre zonas verdes en el periodo 2000 a 2016, éste último como año de análisis del comportamiento espacial de los indicadores.

Para el caso de la dimensión social, se incluyeron los indicadores relacionados con la cercanía de los hogares a vías de tipo 1 y 2 que fueron los de mayor relevancia en el estudio de caracterización de la zona (ver apéndice A); estos indicadores se ubican en el rango intercuartílico 473.53 (18.06;491.58) para la distancia de los hogares a vías tipo 1 y para el indicador de distancia a vías tipo 2 corresponde a 82.42 (5.75;88.17). De otro lado, se incluyeron los casos de enfermedad respiratoria identificados en la zona de estudio con un rango intercuartílico de 5.7 (0.32;6.02). Los indicadores que califican esta dimensión aportan a los subtemas demografía, cobertura de servicios, salud, educación y transporte. No se incluyeron indicadores relacionados con el subtema seguridad dado que la información espacial de los eventos asociados no se encuentra disponible por motivos de seguridad. Por otra parte, es pertinente mencionar que los indicadores en materia de salud no cuentan con una identificación específica a nivel espacial para cada punto en la localidad, por lo cual sólo se incluyó la información recopilada en campo para la caracterización de la zona de estudio y su relación con la calidad del aire (ver apéndice A).

Tabla 4-8 Indicadores en la dimensión social para el análisis espacial

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
Social	Población	Densidad poblacional	(****)	98,002.41	57,107.18	199,296	15,528
	Near_transp	Acceso a sistema de transporte	(****)	66.02	93.05	500	0
	Near_V1	Cercanía a vías T1 (afectación a la salud)	(**)	355.49	422.11	1791.83	1.21
	Near_V2	Cercanía a vías T2 (afectación a la salud)	(**)	83.06	147.85	950.08	0
	Enferm.	Casos de enfermedad respiratoria		4.2	5.21	29.28	0
	Analfabeti	Analfabetismo	(**)	0.84	1.76	10	0
	BL_Ed	Bajo logro educativo	(**)	0.84	1.76	10	0
	Ins_Escola	Inasistencia escolar	(**)	0.42	0.88	5	0
	Rez_Escol	Rezago escolar	(**)	0.42	0.88	5	0
	BASCPI	Barreras de acceso a servicios de cuidado de la primera infancia		0.42	0.88	5	0
	Trab_Inf	Trabajo infantil	(**)	0.42	0.88	5	0
	BASN	Barreras de acceso a salud dada una necesidad	(**)	0.84	1.76	10	0

Subtema	Cod	Indicador		Análisis de los indicadores			
				Media	Desviación estándar	Max	Min
	SAFMA	Sin acceso a fuente mejorada de agua	(**)	0.33	0.71	4	0
	IE_Exc	Inadecuada eliminación de excretas	(*)	0.33	0.71	4	0
	MI_Pis	Material inadecuado de pisos	(*)	0.33	0.71	4	0
	MI_Pared	Material inadecuado de paredes	(*)	0.33	0.71	4	0
	H_crítico	Hacinamiento crítico	(*)	0.33	0.71	4	0
Dimensión económica; interacción habitable (*), equitativa (**), sostenible (****). Los indicadores no marcados no hacen parte de las interacciones y pertenecen puntualmente a la dimensión social. Los ODS relacionados son 1 al 6, 9, 11 y 16							

Al realizar un análisis de correlación de Pearson se encontró correlación positiva lineal entre los indicadores PM_{10}, PM_{25} (0.61); cercanía a transporte y cercanía a vías (0.61); espacios verdes (NDVI) y temperatura superficial terrestre (TST) (0.45); $PM_{2.5}$ y casos de enfermedad respiratoria (0.21); TST y enfermedad respiratoria (0.22). Se encontró además relaciones lineales negativas entre el NDVI y densidad poblacional (-0.27), PM_{10} y densidad poblacional (-0.5), en todos los casos con valores de significancia inferiores a 0.05.

Aunque las correlaciones lineales en el análisis se presentan como débiles, si existe un grado de relacionamiento entre las variables. Para el caso de la concentración de contaminantes atmosféricos y densidad poblacional la relación tiene sentido al verificar las zonas que a nivel espacial presentan mayor concentración de los contaminantes, así como las actividades y usos del suelo que persisten en dichas zonas. Similar situación se refleja en el estudio desarrollado por (Gómez-Losada et al., 2019) al analizar la contaminación de fondo en otras zonas urbanas. De otro lado, la correlación de Pearson para los indicadores generados a partir del índice de pobreza multidimensional presenta correlación perfecta, dado que se partió de una base de cálculo compartida.

El nivel de sostenibilidad para cada zona se presenta en la figura 4-10. De acuerdo con la información que se registra para el año 2016 se encuentra que la zona Norte y Oriente de la localidad presentan un mejor comportamiento en cuanto a la evaluación de la sostenibilidad, con valores máximos de 0.82. Los niveles de sostenibilidad en la zona Sur y Occidente presentan una evaluación en el nivel medio, con valores de evaluación cercanos a 0.68. También se presentan algunas zonas con niveles de sostenibilidad en el nivel bajo con valores mínimos cercanos a 0.21; estas son algunas zonas de borde en el norte de la localidad y en el oriente y al sur en la UPZ Carvajal.

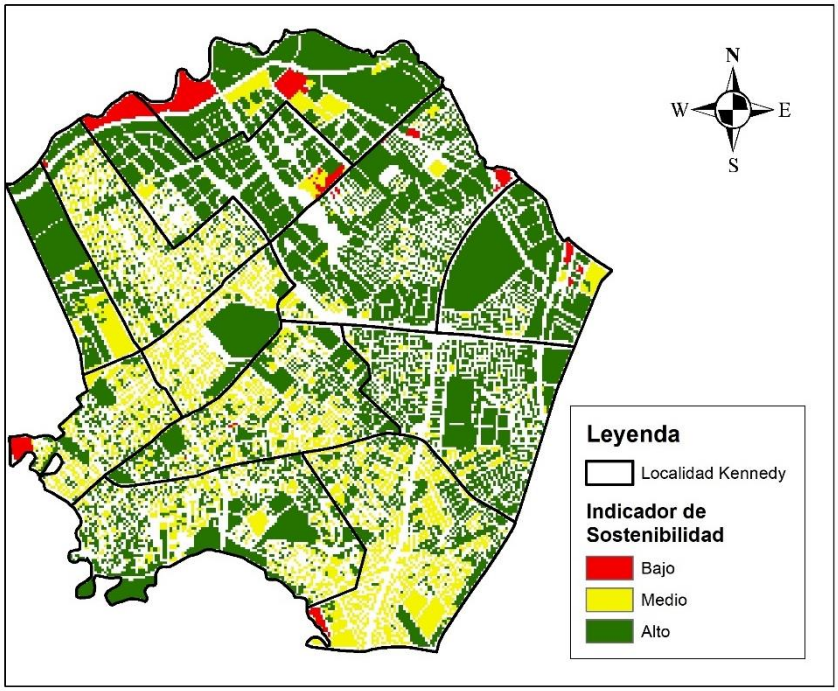


Figura 4-10 Comportamiento espacial del nivel de sostenibilidad para la localidad de Kennedy

Según el comportamiento de los indicadores para el nivel territorial se encuentra que a nivel ambiental, la mayor presión en relación con la calidad de los recursos e influencia en la población, se presenta en las zona sur occidental y occidental de la localidad; esto se contrasta en el comportamiento del material particulado en la zona sur occidental, que es la que ha registrado la mayor concentración del contaminante; además, las zonas con mayores dificultades en la dimensión económica y social coinciden con dichos sectores.

Los resultados que se muestran reflejan de cierto modo la inequidad en la distribución de las cargas y beneficios ambientales, pues las fuerzas impulsoras de actividades económicas que allí se presentan no concluyen en el mismo territorio; es decir, las actividades económicas que allí se efectúan, incluida la operación del sistema de transporte no sólo beneficia al mismo sector, genera presiones con beneficio para otras zonas de la ciudad en la que se ubica la localidad, la región y otros límites territoriales.

Existe un desequilibrio entre el factor económico que impulsa la generación de presiones ambientales en el territorio y que requiere un análisis de mayor amplitud desde la articulación de las presiones en la zona de origen y los productos en la zona de destino, para evidenciar las cargas ambientales, económicas y sociales en cada una de las localidades de la ciudad de Bogotá. Las UPZ Patio Bonito, Corabastos y Carvajal reflejan importantes cargas de orden ambiental, alta densidad poblacional y condiciones en la dimensión social que contradicen el concepto de justicia ambiental, conocido como la distribución equilibrada de cargas y beneficios ambientales entre todas las personas en el territorio analizado, bajo el reconocimiento de las condiciones comunitarias y su participación en la adopción de decisiones que le afectan (Espejo, 2010).

4.2 Herramientas de aprendizaje automático aplicables en el contexto del desarrollo sostenible y la calidad del aire

Como se mencionó con anterioridad, el desarrollo sostenible corresponde con la integración de temas y subtemas en el marco de las dimensiones ambiental, social, económica e institucional, según la Comisión de Desarrollo Sostenible de las Naciones Unidas. En la dimensión ambiental se encuentran los siguientes grandes temas: atmósfera, océanos costas y mares, suelo, agua dulce, biodiversidad y que incluyen subtemas de gran relevancia en el contexto urbano como lo son la calidad del aire, cantidad y calidad del agua y ecosistemas. Para la dimensión social los grandes temas son: educación, salud, trabajo, pobreza, seguridad, vivienda y para la dimensión económica: estructura económica y patrones de producción y consumo. Finalmente, la dimensión institucional abarca temáticas asociadas a la política pública y gobierno, que incluyen además a la capacidad institucional.

Para todos los temas y subtemas en las diferentes dimensiones se incluyen indicadores que permiten identificar su comportamiento para soportar decisiones, principalmente en el marco político. Además, tanto los temas, subtemas, objetivos, metas e indicadores estarán determinados por las condiciones específicas del territorio.

En la identificación de las experiencias existentes, relacionadas al pronóstico del desarrollo o desempeño sostenible de los territorios, se encontraron pocos estudios (ver figura 4-11 y en mayor detalle los resultados descritos en el apéndice C); éstos se orientaron principalmente a la clasificación y ranking de niveles de sostenibilidad entre naciones y a la evaluación de niveles de sostenibilidad de sectores económicos específicos.

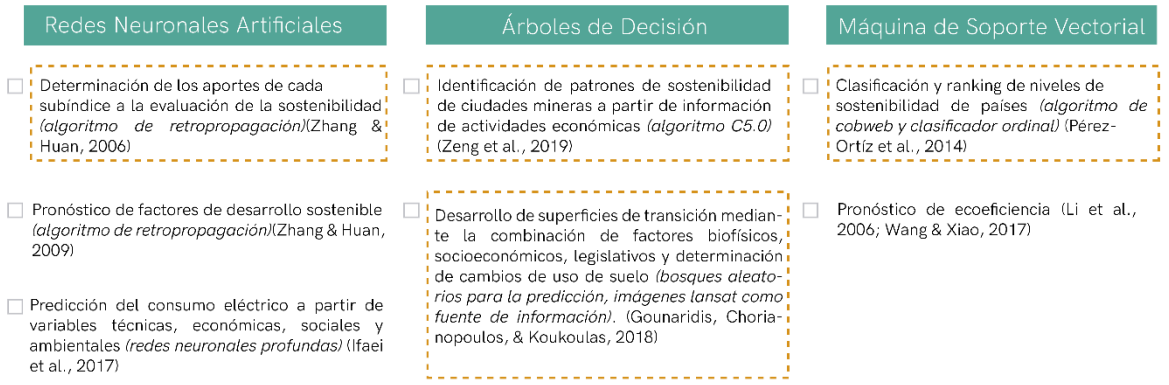


Figura 4-11 Métodos y estudios aplicados en el ámbito global para la predicción de la sostenibilidad y/o dimensiones de la sostenibilidad mediante herramientas de aprendizaje automático

En la mayoría de los estudios se aplicaron herramientas de aprendizaje automático para generar un indicador temático (Gounaridis et al., 2018; Ifaei et al., 2017), abordando el comportamiento de factores económicos, ambientales y sociales (Pérez-Ortíz et al., 2014; Zeng et al., 2019; Zhang et al., 2009; Zhang y Huan, 2006), para establecer consumos energéticos y otros orientados a la predicción de la ecoeficiencia (Li et al., 2006; Wang y Xiao, 2017). La mayoría de los estudios aplicaron las redes neuronales artificiales (ANN, por sus siglas en inglés), seguidos de la máquina de soporte vectorial (SVM, por sus siglas en

inglés) y los árboles de decisión (DT por sus siglas en inglés). Los estudios avanzaron con la aplicación de ANN para ponderar la importancia de los factores en la integración de las dimensiones, y otros utilizaron SVM y DT para la predicción a partir de patrones en la integración de las dimensiones de la sostenibilidad. Mayor detalle se presenta en el apéndice C.

De las experiencias existentes se destacan varios elementos: en primer lugar, 1) el algoritmo de retro propagación fue el algoritmo más utilizado en los trabajos que aplicaron redes neuronales; en segundo lugar, 2) una gran ventaja de los estudios que compararon el comportamiento de las naciones y sectores es que se cuenta con información en cada sector o territorio analizado, a pesar de trabajar con información anual en los ámbitos de estudio, existe un volumen de información útil para su procesamiento; en tercer lugar 3) aunque los estudios coinciden en el uso de algunos indicadores, se encuentra variedad en su elección y número de factores que componen a las dimensiones de sostenibilidad a analizar; seguidamente, 4) el uso de imágenes satelitales e información espacial es un insumo relevante para el análisis de comportamiento espacial de los territorios y por último 5) el acercamiento al tratamiento espacial de los niveles de sostenibilidad se ha desarrollado a través de la aplicación de algoritmos de agrupamiento como el de k vecinos más próximos (k-nn, por sus siglas en inglés). Las experiencias establecidas en dichos estudios orientaron la elección de los métodos de aprendizaje automático, los algoritmos, así como instrumentos de recopilación de información; estas experiencias se destacan en la figura 4-11 en los recuadros de color naranja.

En relación con el subtema calidad del aire, que es objeto de especial interés en esta investigación y, teniendo en cuenta que la mayoría de los estudios se enfocan en subtemas y temas específicos en la dimensión ambiental, se encontró la aplicación de las redes neuronales artificiales, la máquina de soporte vectorial y los árboles de decisión para el pronóstico de contaminantes que influencian la calidad del aire en diversas escalas espaciales y temporales.

La mayoría de los estudios coincide en el uso de un conjunto de datos de entrenamiento en una proporción entre 60 a 80 % de los datos y la proporción restante, para la validación de los modelos de pronóstico. Del mismo modo coinciden en que el material particulado en fracciones inferiores a 10 micras (PM_{10}), incluidas las menores a 2.5 micras ($PM_{2.5}$) son los contaminantes más frecuentes en los estudios de pronóstico (mayor detalle de esta información se encuentra en el apéndice C). Además, los estudios que analizaron la influencia de la calidad del aire en la salud de la población se han basado en modelos epidemiológicos que establecen una relación causal en términos de tiempo de exposición a contaminantes

atmosféricos, dejando de lado la zonificación de puntos de interés espacial al respecto de la exposición y los contaminantes.

La elección de los métodos, de los algoritmos aplicados en este trabajo y la elección de las métricas de desempeño de los modelos se soporta en el trabajo de investigación publicado bajo la siguiente referencia:

Molina-Gómez, N.I., Díaz-Arévalo, J.L. & López-Jiménez, P.A. Air quality and urban sustainable development: the application of machine learning tools. *Int. J. Environ. Sci. Technol.*18, 1029-1046 (2021). https://doi /10.1007/s13762-020-02896-6

4.3 Predicción a partir del uso de herramientas de aprendizaje automático

4.3.1 Predicción con información anual-mensual

Se realizó la predicción utilizando las herramientas de aprendizaje automático árboles de decisión, máquina de vector soporte y redes neuronales artificiales. En este análisis se destaca la temporalidad mensual de la información utilizada para la alimentación del modelo. Es preciso recordar que el uso de información con escala inferior a la anual puede presentar estacionalidad. Sin embargo, en el caso analizado, las etiquetas de clasificación generadas en el cálculo del nivel de desempeño sostenible se encuentran dentro de los rangos de comportamiento anual. Además, una de las ventajas de los clasificadores discriminantes utilizados es que se trata de métodos robustos a la presencia de valores atípicos que pueden ser gestionados como objetos específicos (Gibert et al., 2018) .

Las métricas de desempeño alcanzadas en cada modelo fueron las siguientes (ver tabla 4-9).

Tabla 4-9 Métricas de desempeño de los modelos de clasificación de la sostenibilidad con información temporal (Molina-Gómez et al., 2020)

Modelo	**Exactitud balanceada**			**Precisión**			**Sensibilidad**			**Especificidad**		
	alto	*medio*	*bajo*	*alto*	*medio*	*bajo*	*alto*	*medio*	*bajo*	*alto*	*medio*	*bajo*
Árboles de desición (C-5.0Tree)	0.95	0.81	0.6	0.75	0.90	0.33	1.0	0.82	0.33	0.91	0.80	0.86
Redes neuronales artificiales -Nnet	0.96	0.8	0.5	0.75	0.85	-	1.0	1.0	0.00	0.92	0.60	1.0
Máquina de soporte vectorial - SVMradial	0.79	0.7	0.5	0.67	0.79	-	0.67	1.0	0.00	0.92	0.40	1.0

Como se presenta en la figura 4-12 la clasificación del nivel alto establece las mejores métricas de desempeño para los tres modelos aplicados, en donde los árboles de decisión y las redes neuronales artificiales presentan los mejores resultados; situación similar ocurre en la clasificación del nivel medio de sostenibilidad. Para el nivel bajo, sólo el modelo de árboles de decisión presenta resultados completos de las métricas de evaluación; lo cual se explica en que el set de datos para esta etiqueta no posee un gran volumen de información.

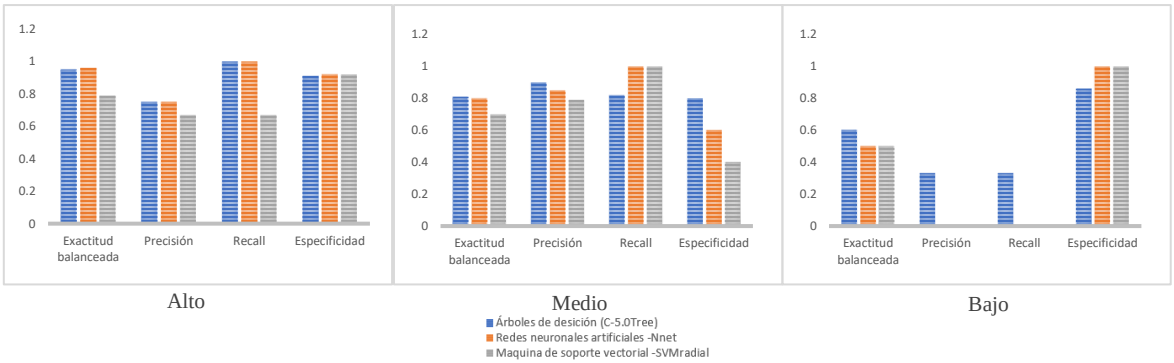


Figura 4-12 Métricas de desempeño de los modelos en la clasificación de los niveles de sostenibilidad

Por otra parte, la Agenda 2030, sus metas, indicadores y los planes de gobierno ajustados para alcanzar el logro de los ODS establecen que el valor esperado en el nivel de desempeño de los territorios tienda a la mejora; es decir, pasar de una etiqueta de clasificación "bajo" a "medio" y superar de manera sostenida esta categoría llegando a un nivel alto. En este orden de ideas, es posible que la etiqueta de clasificación "bajo" tienda a desaparecer, y la clasificación se centre en el nivel medio y alto, pues con la mejora de las condiciones del territorio la categoría bajo no tendría información disponible para la alimentación del modelo de predicción. No obstante, en el tiempo, algunas condiciones relacionadas a eventos inesperados, que modifican el comportamiento convencional de las dimensiones de la sostenibilidad podrían hacer que la categoría "bajo" incluyera nueva información y la clasificación establecida se modificara.

4.3.1.1 Variables de importancia en la modelación con escala temporal

La clasificación de los niveles de sostenibilidad está influenciada por las variables explicativas, algunas con mayor influencia en la clasificación que otras. Por lo tanto, a partir del procedimiento descrito en el epígrafe 3.5 se identificó la relevancia de las variables en el proceso de clasificación.

La variable predictora de mayor incidencia fue el acceso a servicios de salud (S47), seguido de reportes de violencia y abuso doméstico, familiar e infantil (S32) y pasajeros transportados por el sistema de transporte público masivo (S35); también es relevante la variable energía consumida en Kwh/año (EC13). En los tres modelos la variable predictora denominada excedencias a la norma de calidad del aire para PM_{10} (E3) presenta una relevancia superior al 50%. El modelo de árboles de decisión tuvo tres variables predominantes en su proceso de clasificación S47, E3 y el consumo percápita de agua residencial (EC16), mientras que en el modelo de redes neuronales todas las variables presentan algún grado de participación. En la figura 4-13 se presentan el nivel de importancia de las variables.

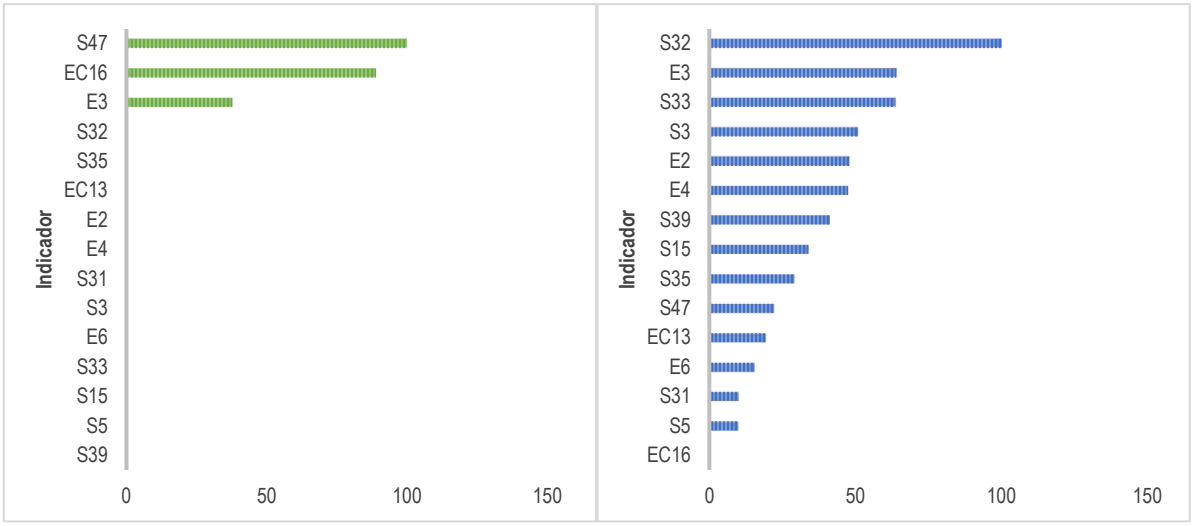


Figura 4-13 Variables de importancia según modelo de clasificación con árboles de decisión y con redes neuronales artificiales para los niveles alto, medio y bajo de sostenibilidad

En relación con el modelo en el que se aplicó ANN, la variable S32 reportó la mayor influencia, seguido de E3 con el 64%, el robo simple y agravado (S33) y la tasa de mortalidad por neumonía con el 50.9%. La variable $PM_{2.5}$ presentan un factor de influencia del 47.9%. Contrario al modelo en el que implementó SVM la variable EC16 no presenta ninguna influencia. Para el modelo de SVM (ver figura 4-14) las variables relacionadas con la calidad del aire E3 y E2 presentan una influencia cercana al 88% para la clasificación del avance en el desarrollo sostenible en los niveles medio y alto; para el caso de la clasificación en nivel bajo la variable E2 presenta una importancia cercana al 59%.

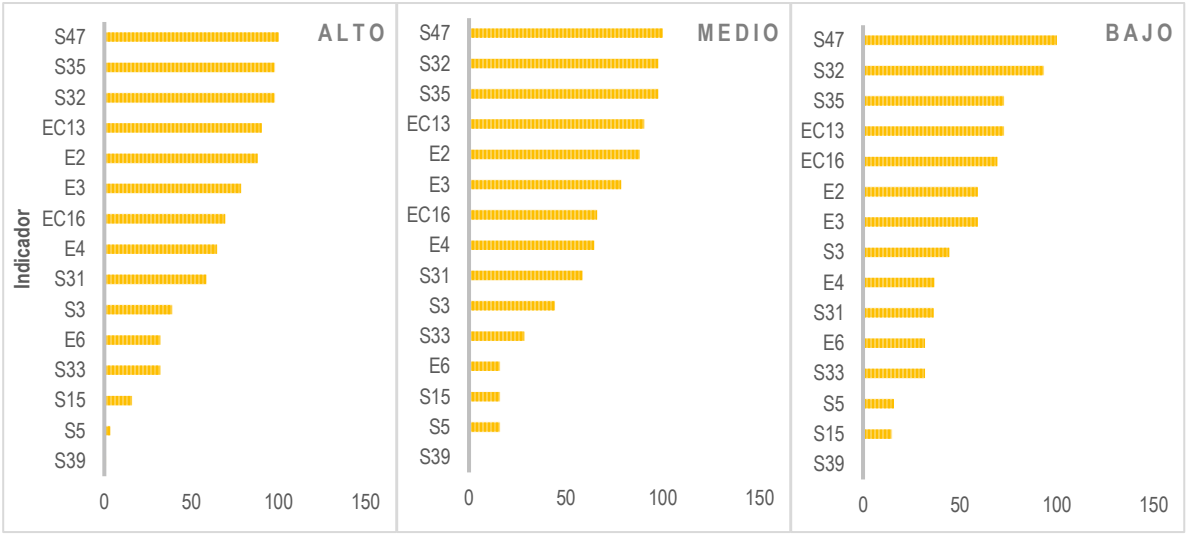


Figura 4-14 Variables de importancia según modelo de clasificación con la máquina de vector soporte para los niveles alto, medio y bajo de sostenibilidad

4.3.2 Predicción en el ámbito espacial

A partir de la composición de bandas con la información espacial de las variables explicativas (ver figura 4-15) y las etiquetas de clasificación del nivel de sostenibilidad alto, medio y bajo, se llevó a cabo el proceso de predicción. Las herramientas de aprendizaje automático utilizadas para ello fueron los bosques aleatorios, las redes neuronales artificiales y la máquina de vector soporte.

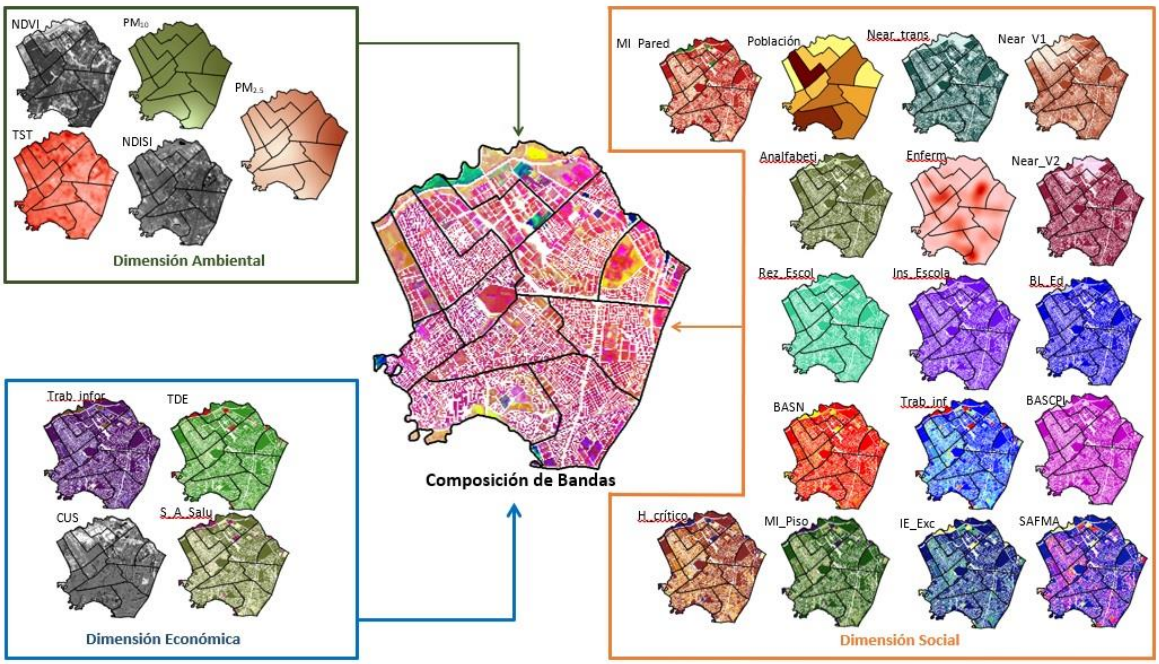


Figura 4-15 Composición de bandas utilizada para la predicción espacial

Las métricas de desempeño alcanzadas por cada modelo y en cada predicción según la etiqueta de clasificación se detallan en la tabla 4-10.

Tabla 4-10 Métricas de desempeño de los modelos evaluados en el proceso de clasificación

Modelo	**Exactitud balanceada**			**Precisión**			**Sensibilidad**			**Especificidad**		
	alto	*medio*	*bajo*	*alto*	*medio*	*bajo*	*alto*	*medio*	*bajo*	*alto*	*medio*	*bajo*
Bosque aleatorios	0.929	0.920	0.983	0.962	0.889	0.968	0.967	0.872	0.968	0.891	0.968	0.998
Redes neuronales artificiales -Nnet	0.931	0.923	0.970	0.963	0.887	0.976	0.967	0.878	0.941	0.894	0.967	0.999
Máquina de soporte vectorial -SVMradial	0.922	0.911	0.976	0.957	0.892	0.981	0.971	0.853	0.953	0.874	0.970	0.999

Los tres modelos aplicados generaron métricas de desempeño destacables, los modelos de bosques aleatorios y redes neuronales presentan mayor similitud en cuanto a los resultados de las métricas. Para el caso de la exactitud balanceada, el modelo de redes neuronales presenta la mayor exactitud en la clasificación de las etiquetas de nivel de sostenibilidad en los niveles alto y medio; en tanto que el modelo de bosques aleatorios se destaca con una exactitud balanceada del 98.3% en la clasificación del nivel de sostenibilidad bajo. Sin

embargo, se destaca que los resultados del modelo de máquina de soporte vectorial no son distantes al desempeño ofrecido en los demás modelos.

El uso de la métrica exactitud balanceada obedece a que, a pesar de garantizar la selección aleatoria de la información para el entrenamiento del modelo, el comportamiento del territorio en sus diferentes indicadores no es homogéneo, por lo que la concentración de condiciones menos favorables puede evidenciarse con más claridad en unas zonas respecto a otras (ver figura 4-15).

La métrica denominada precisión (ver tabla 3-2) relaciona los resultados de clasificación apropiadamente predichos (verdaderos positivos) respecto de la sumatoria de las etiquetas erróneamente clasificadas como verdaderas según cada nivel de sostenibilidad (falsos positivos) y la etiqueta objetivo apropiadamente clasificada (verdaderos positivos). Esta métrica registró para la clasificación del nivel alto el mejor comportamiento en el modelo de redes neuronales (96.3%). Para el caso de los niveles medio y bajo fue la máquina de soporte vectorial la que generó las mejores métricas de clasificación con el 89.2% y 98.1% respectivamente (ver tabla 4-10).

En el caso de la métrica sensibilidad, exhaustividad o tasa positiva real que corresponde a la fracción de instancias relevantes (verdaderos positivos) recuperadas del conjunto de información de la misma clase disponible para la clasificación (ver tabla 3-2), el nivel alto presentó las mejores métricas en el modelo de SVM (97.1%), mientras que el nivel medio presentó las mejores métricas en el modelo de ANN (87.8%) y el nivel bajo en el modelo de RF (96.8%).

La media armónica entre precisión y sensibilidad, correspondientes al puntaje F1, establece que el modelo con el mejor desempeño fue el modelo de redes neuronales artificiales (96.5% y 88.2% para la clasificación de los niveles alto y medio), seguido del modelo de bosques aleatorios (96.8% para la clasificación del nivel bajo). Al tratarse de una muestra desbalanceada en lo relacionado con la clasificación del nivel de sostenibilidad, la media armónica de la medida de F1 es el instrumento más apropiado para la definición del mejor modelo de clasificación.

4.3.2.1 Variables de importancia en el modelo

Las variables relacionadas con la contaminación del aire (PM_{10}, $PM_{2.5}$), expansión urbana y zonas impermeables fueron las variables de mayor importancia en el modelo de redes neuronales, seguido de variables socioeconómicas. Para el caso del modelo de bosques aleatorios se observa una marcada importancia de las variables PM_{10}, seguido por las variables socioeconómicas: trabajo infantil, material de las paredes exteriores y trabajo informal y $PM_{2.5}$. Para este modelo la cercanía a las vías es una variable de menor importancia en el modelo de clasificación de los niveles de sostenibilidad (ver figura 4-16).

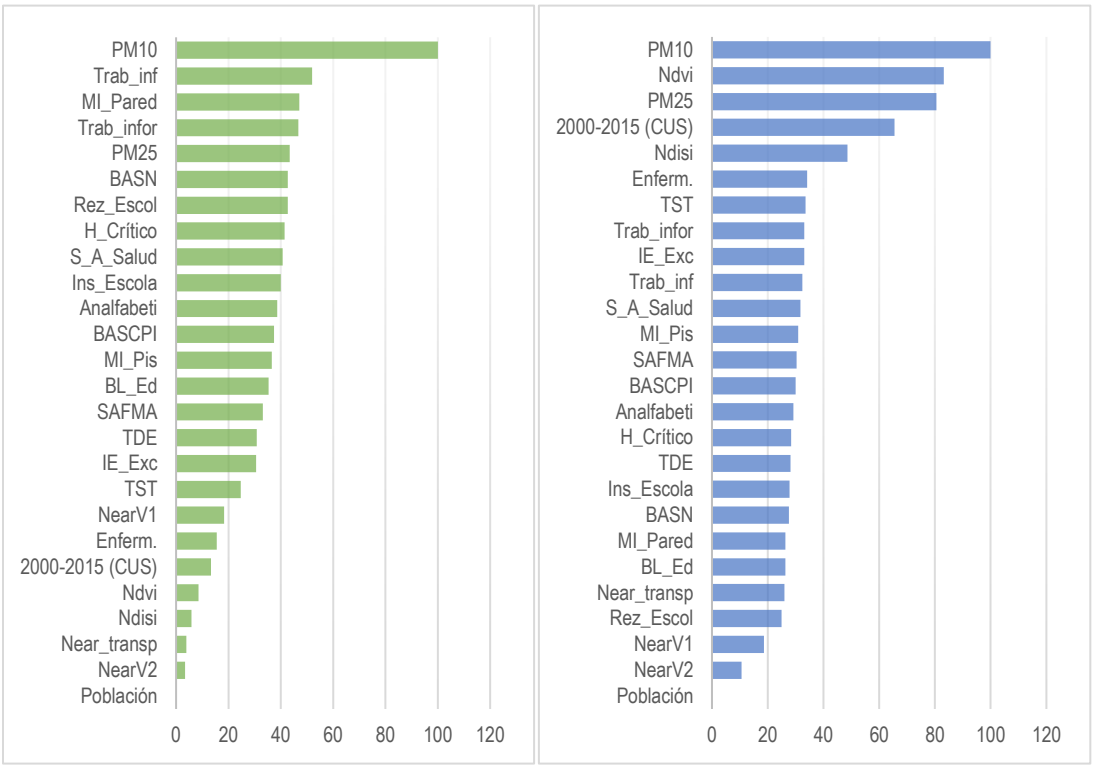


Figura 4-16 Variables de importancia según modelo de clasificación bosques aleatorios (barras verdes) y redes neuronales artificiales (barras azules) para los niveles alto, medio y bajo de sostenibilidad

La calidad del aire es un indicador que es modificado en virtud de las condiciones del territorio, las actividades económicas que se desarrollen, las fuentes de emisión y características de las emisiones. Además, su calidad está fijada por las condiciones específicas del territorio en lo relacionado con la dispersión de los contaminantes en la atmósfera; para zonas urbanas se presentan cañones urbanos dada la ubicación de edificaciones, que con el paso de los años han incrementado en altura y se ha presentado una expansión modificando la disponibilidad de coberturas vegetales óptimas para la retención de material particulado y también para aportar a los fines de mitigación del fenómeno de isla de calor urbano.

La calidad del aire es un determinante intermedio que influencia la salud de la población, contribuyendo con incrementos en los costos sociales asociados a las tasas de morbilidad y mortalidad y el pago de tratamientos relacionados. Los indicadores elegidos en este trabajo corresponden al material particulado como principal contaminante y el más reconocido y analizado a nivel mundial, éste ha sido uno de los elementos más predominantes en el proceso de clasificación del nivel de sostenibilidad en la escala temporal como espacial. A partir del uso de herramientas de aprendizaje automático se comprueba la influencia de los contaminantes atmosféricos en el propósito de clasificación.

4.4 Propuesta metodológica para el análisis de la influencia de la calidad del aire en el Desarrollo Urbano Sostenible, a partir de aprendizaje automático

Además de los resultados anteriormente descritos, se estableció como producto final una propuesta metodológica para la determinación de la influencia de la calidad del aire en el desarrollo urbano sostenible y cuyas etapas se presentan en la figura (4-17). El procedimiento seguido y los resultados generados, según el desarrollo de cada objetivo específico planteado, permitieron concretar en una metodología que inicia con la elección de indicadores y finaliza con la identificación de las variables que influencian el modelo de predicción de los niveles de sostenibilidad.

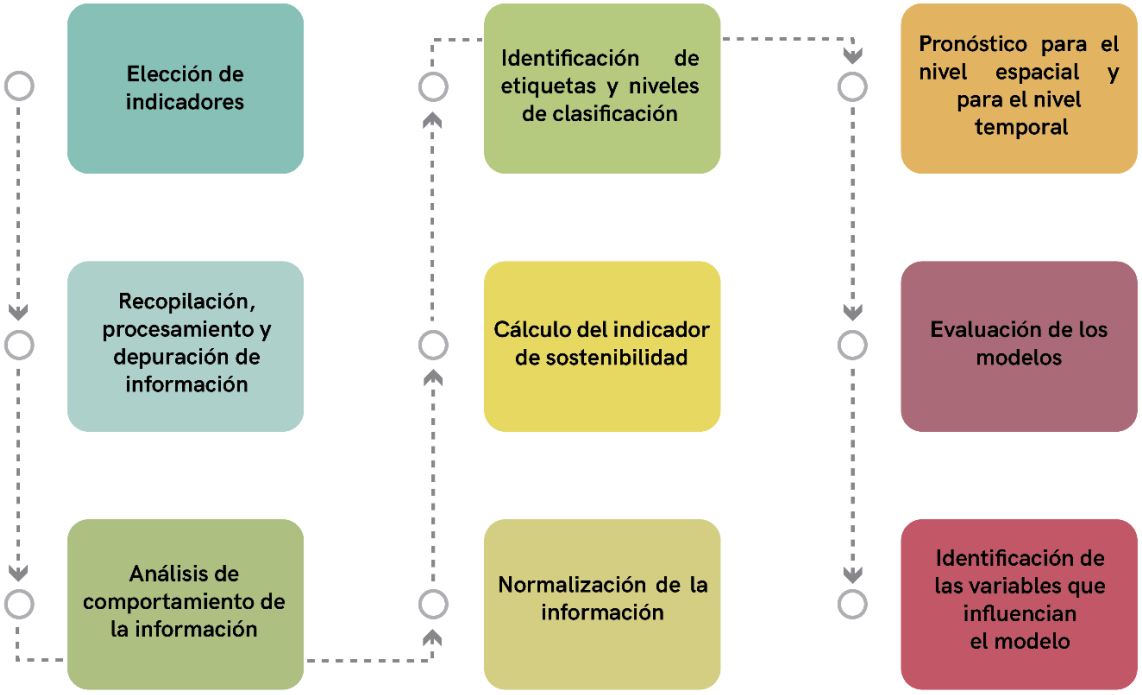


Figura 4-17 Propuesta metodológica para la determinación de la incidencia de la calidad del aire en el desarrollo urbano sostenible (Adaptado de Molina-Gómez, et al., 2020)

En la ejecución de cada etapa fue posible establecer las mejores prácticas requeridas para la aplicación metodológica y que se resumen en la tabla 4-11, con la identificación de las etapas, su descripción a través de la definición de los elementos centrales de su implementación, los insumos e información requerida.

Tabla 4-11 Descripción metodológica propuesta para la identificación del nivel de influencia de la calidad del aire en el desarrollo urbano sostenible

Etapa	Elementos centrales de la etapa	Insumos/información requerida
1. Elección del conjunto de indicadores que mejor califica a la zona de estudio	-Elección del marco de trabajo, periodo y territorio objeto de evaluación y análisis -Definición de las necesidades de la comunidad a partir de: instrumentos como quejas y peticiones, el análisis de información en medios de comunicación que	1. Marco de indicadores de las naciones unidas 2. Reporte de quejas y peticiones de la comunidad en el territorio

Etapa	Elementos centrales de la etapa	Insumos/información requerida
	se allega a la comunidad y resultados de trabajo con grupos de interés -Definición de variables de calificación de los indicadores -Elección de indicadores, involucrando metas y lineamientos de política -Clasificación de indicadores según dimensión, tema, subtema y aporte a los ODS	3. Resultados de talleres con la comunidad para identificar requerimientos 4. Información digital de medios de comunicación
2. Recopilación, procesamiento y depuración de la información	- Recopilación de información en la menor unidad temporal disponible para todos los indicadores - Recopilación de información espacial para todos los indicadores - Análisis y eliminación de valores atípicos	Bases de información, información en formato *shape* disponible en observatorios o sitios de datos abiertos y respuestas oficiales de entidades, imágenes satelitales, índices espectrales generados a través del análisis de imágenes satelitales
3. Análisis de comportamiento de la información	-Aplicación de estadísticas descriptivas para el conjunto de información en las escalas temporal y espacial -Análisis de correlación de las variables y aplicación de análisis de correlación canónico para determinar posibles relaciones entre los indicadores en las dimensiones e interacciones (se requiere normalización de los datos)	Información temporal y espacial de la etapa 2, según indicadores elegidos en la etapa 1
4. Normalización de la información	-Aplicación del método de normalización llevando los valores de los indicadores a valores entre 0 y 1, (ecuación 3, tabla 3-1) -Organización de la información de tal manera que los valores cercanos a 0 registren las condiciones de menor calidad o desempeño del indicador y los valores cercanos a 1 la mejor calidad o desempeño del indicador, -Realizar comparación de los valores registrados para cada indicador con los valores límite/umbrales establecidos en regulaciones, políticas, metas y objetivos específicos en cada caso	Indicadores categorizados de acuerdo con la dimensión a la que pertenece
5. Cálculo del indicador de sostenibilidad	Mediante el uso de las ecuaciones 1, 2, 4 (ver tabla 3-1) y 5 (ver tabla 3-3) se realiza el cálculo a nivel espacial y para cada periodo elegido en la evaluación del nivel de avance en el desarrollo sostenible	Indicadores normalizados
6. Identificación de etiquetas y niveles de clasificación	-De acuerdo con los rangos de clasificación establecidos (alto, medio, bajo) se define el nivel de clasificación para cada periodo de análisis (evaluación temporal) y en el territorio en su conjunto (evaluación espacial).	1. Subíndices para cada dimensión 2. Índice de sostenibilidad para cada año y a nivel espacial
7. Pronóstico para el nivel espacial y para el nivel temporal	Se ingresa la información en el modelo de clasificación espacial y en el modelo de clasificación temporal. Para el primer caso se	1. Información espacial de los diferentes indicadores según dimensión y agregados en

Etapa	Elementos centrales de la etapa	Insumos/información requerida
	hace uso del modelo de ANN ó RF con los parámetros de calibración establecidos en el numeral 3.4.2.2. Para el caso de la clasificación temporal se sugiere iniciar con el uso de árboles de decisión de acuerdo con los parámetros de calibración establecidos en el apéndice B; en la medida que se cuente con mayor disponibilidad de información es posible hacer uso de instrumentos de predicción más robustos como las redes neuronales artificiales.	una sola capa de información, mediante una composición de bandas 2. Información temporal de los diferentes indicadores organizada como *data frame* según dimensión 3. Etiquetas de clasificación a nivel espacial y temporal
8. Evaluación de los modelos aplicados a partir de métricas de desempeño	- En caso de contar con muestras desbalanceadas se sugiere aplicar la métrica de exactitud balanceada - Realizar un análisis comparativo de las métricas precisión y sensibilidad determinando la medida F	Matriz de confusión según el pronóstico espacial y el pronóstico temporal
9. Establecer las variables que influencian el modelo de clasificación	- Identificar a partir de la función "importance" para cada clasificador y la medida de Gini las variables de mayor influencia en el proceso de clasificación, así podrá establecer el grado de influencia de indicadores de calidad del aire en el desarrollo urbano sostenible, como otras variables que requieren de atención	Variables ingresadas en el modelo

Para el desarrollo de la etapa 1 es necesario utilizar como herramientas complementarias el análisis de priorización jerárquico (AHP), el método Delphi, el análisis de necesidades de la comunidad (con la comunidad y quejas) y el análisis con minería de texto para identificar temáticas de comunicación a la comunidad. En la etapa 2 se requiere del software ArcGis y en todo el proceso metodológico el uso de lenguajes de programación como R, Python u otros programas requeridos para el procesamiento, análisis de la información, y el pronóstico.

5 Capítulo 5 Conclusiones y Desarrollos Futuros

Este capítulo se construyó como parte del quinto y último objetivo específico planteado para el desarrollo de este trabajo doctoral, que fue alimentándose a partir de las principales conclusiones en cada etapa del proyecto.

5.1 Conclusiones

Dentro de las principales conclusiones generadas en relación con el primer objetivo específico de este trabajo, se destaca que la zona urbana en la que se aplicaron los elementos técnicos de esta investigación refleja características de territorios con alta densidad poblacional; problemas de contaminación del aire, especialmente en el occidente y sur de la localidad; incremento de zonas impermeables en el paso de los años y, la coexistencia de diferentes usos del suelo que permiten que la población se exponga a contaminantes atmosféricos. Dicha exposición incide en términos generales en el bienestar y en la salud de la población.

Se concluye además que, a pesar de la tendencia decreciente en la concentración de contaminantes atmosféricos en la zona, en el periodo 2009-2017, las concentraciones del PM_{10} y el $PM_{2.5}$ superaron los valores guía establecidos por la OMS (PM_{10}= 20 µg/m^3; $PM_{2.5}$= 10 µg/m^3). Bajo estas condiciones es posible un incremento en la mortalidad por enfermedades cardiopulmonares y de cáncer de pulmón. Se destaca que las áreas de mayor interés en la zona urbana, según el pronóstico de zonas con mayor probabilidad de presentar enfermedades del sistema respiratorio, corresponden a espacios en las UPZ Calandaima y Patio Bonito (ubicadas al Noroccidente de la localidad). Esta última es una de las UPZ más densamente pobladas y coincide con una de las zonas en las que se presentan los mayores valores de concentración de $PM_{2.5}$ para el año 2016.

La manifestación de eventos asociados con el sistema respiratorio para este tipo de zonas puede relacionarse además con una exposición crónica a contaminantes atmosféricos dado que, en el trabajo de campo realizado en 2017, el 49.2% de la población manifestó haber permanecido en la localidad por más de diez años. Además, en el amplio espectro de enfermedades del sistema respiratorio (desde resfriado común hasta infecciones agudas del sistema respiratorio), el 21.4% de la población manifestó que alguno de los miembros de su hogar fue diagnosticado en 2016 con alguna enfermedad del sistema respiratorio, de esta cifra el 51.8% corresponde a individuos en edad de trabajar y 14.3% a individuos con edad superior a los 60 años.

Adicionalmente, siendo el modelo de bosques aleatorios el que presentó las mejores métricas de desempeño (AUC= 0.63 y exactitud =77.5%) para el pronóstico de zonas con mayor probabilidad de presentar enfermedades del sistema respiratorio, se encontró que el $PM_{2.5}$ y la proximidad de los hogares a vías primarias (T1) y secundarias (T2), fueron las variables de mayor influencia en el proceso de pronóstico. Esto se explica en las condiciones del tráfico y emisiones propias de vehículos de carga y de transporte de pasajeros; estos aspectos se

relacionan con la densidad poblacional, así como la cercanía a la central de abastecimiento de alimentos de la ciudad capital, en la UPZ Corabastos. A partir del análisis de correlación de Pearson, se encontraron relaciones lineales positivas entre contaminantes atmosféricos, la densidad poblacional y el $PM_{2.5}$.

Puede concluirse además que, considerando la importancia de evaluar el nivel de influencia de la calidad el aire en el desarrollo urbano sostenible, se identificó un conjunto de indicadores para el análisis temporal y espacial del desempeño del territorio. En total se construyó una matriz de 81 indicadores con información anual de la localidad (en el periodo 2009-2017) y 26 indicadores para el análisis del comportamiento espacial. El número de indicadores se redujo en el análisis espacial, dado que la disponibilidad de información es limitada en ambas escalas; principalmente para el análisis de micro territorios.

La dimensión social registró el conjunto de mayor número de indicadores (58% para el análisis temporal y 65% en el análisis espacial), dadas las condiciones sociales que se requieren para el apropiado desempeño de una zona urbana; la dimensión económica (19% para el análisis temporal y 15% en el análisis espacial) incluyó indicadores que analizan principalmente, las condiciones de pobreza, empleo e ingresos de la población y la dimensión ambiental (17% en el análisis temporal y 19% en el análisis espacial), no sólo incluyó indicadores que califican las condiciones de calidad del aire, sino también condiciones ambientales del territorio urbano.

La calidad del aire se ve influenciada por elementos como el transporte como fuente móvil de emisión, zonas y espacios verdes, usos del suelo, densidad poblacional, el comportamiento de variables meteorológicas, incluida la temperatura superficial terrestre. Todos estos elementos se incluyeron como indicadores en el análisis de las dimensiones del desempeño sostenible, cuyo fundamento partió del reconocimiento del micro territorio en la relación entre calidad del aire y salud de la población.

De los elementos analizados para la identificación de variables que podrían calificar las dimensiones individuales del desempeño sostenible, se encontró que la comunidad presenta solicitudes en la dimensión social relacionados con violencia, seguridad, salud, educación y recreación, principalmente. Temas que hacen que la dimensión social deba presentar un conjunto de indicadores variado para analizar el comportamiento en virtud de cada requerimiento. Adicionalmente, respecto a la información divulgada a la comunidad en medios de prensa digital, se encontró que existe poca cohesión con los fines globales del desarrollo sostenible, siendo frecuente la entrega de información puntualizada y de baja articulación con los ámbitos de sostenibilidad; no obstante, los temas recurrentes informados a la comunidad corresponden a elementos clave de los ODS como pobreza, agua, desigualdad, desempleo y violencia.

Puede concluirse además que el marco de trabajo definido por las naciones unidas, a propósito de los ODS, orientó la elección de los indicadores para el análisis de cada dimensión de la sostenibilidad urbana. Este contexto, junto con la revisión de experiencias de otros estudios permitieron establecer las bases para la elección de los indicadores más apropiados, entre los cuales no sólo se incluye la consulta a expertos técnicos, sino también

las necesidades de la comunidad en el área de estudio. Este último elemento, junto al contexto legal, político y de ordenamiento del micro territorio, es necesario a fin de evaluar de la manera más cercana el espacio territorial seleccionado.

Con relación al segundo objetivo específico, se puede concluir que los indicadores que describen las diferentes dimensiones de la sostenibilidad presentan un comportamiento no lineal, por lo que se realizó un análisis de correlación canónica encontrando asociaciones de indicadores en las interacciones habitable, viable, equitativo y sostenible y que reflejan una imagen completa del territorio. En el análisis de los indicadores en la escala temporal se encontró algunos que reflejan condiciones de estabilidad que no mejora o varían con los años y que pueden comprometer las condiciones de habitabilidad del territorio, tal es el caso del tratamiento de aguas residuales, áreas de espacios verdes y de recreación.

Se realizó el cálculo del desempeño sostenible identificando los niveles de avance para cada periodo de análisis y en el ámbito espacial. Para el año 2016 se encontró en el territorio un nivel mínimo de sostenibilidad de 0.21, máximo de 0.82 y promedio de 0.68, con una desviación estándar de 0.09. Lo anterior establece un comportamiento influenciado por los diferentes indicadores en cada dimensión a nivel espacial. Para ese año, el indicador de sostenibilidad general correspondió a 0.64 en el nivel medio de la clasificación, muy cercano a la valoración realizada en el ámbito espacial. Se destaca además que para los años 2009 y 2010 la localidad registró niveles bajos de sostenibilidad y sólo en 2016 se superó el rango de sostenibilidad medio.

Con relación al cumplimiento del tercer objetivo específico se encontraron pocas experiencias en las que se aplicó aprendizaje automático al pronóstico del desarrollo sostenible; la máquina de vector soporte fue la herramienta más utilizada y en menor medida las redes neuronales artificiales, los árboles de decisión y bosques aleatorios. La mayoría de los estudios en el campo del desarrollo sostenible utilizaron la regresión como factor clave en el proceso de pronóstico.

Para el caso de las experiencias con redes neuronales el algoritmo de retro propagación fue el algoritmo más utilizado. El algoritmo C5.0 se utilizó para la identificación de patrones de sostenibilidad con árboles de decisión y, los bosques aleatorios fueron aplicados para la predicción de superficies de transición en la combinación de diferentes factores. SVM se utilizó para la clasificación y ranking de países. La mayoría de los estudios coincide en el entrenamiento del conjunto de datos en una proporción entre 60 a 80 % de la información; con una gran ventaja asociada a la disponibilidad de información útil para su procesamiento respecto de cada sector o territorio analizado.

Finalmente, para el cuarto objetivo específico, consistente en la aplicación de las herramientas de aprendizaje automático seleccionadas, se puede concluir que, con el uso de herramientas como los bosques aleatorios, las redes neuronales y la máquina de vector soporte fue posible predecir el comportamiento espacial del desempeño urbano sostenible en el micro territorio. El pronóstico realizado es consistente con los resultados generados en la clasificación para cada año en el periodo de estudio, en especial el 2016 que fue el año seleccionado para la predicción espacial.

El modelo que mejor desempeño presentó en el análisis de la media armónica entre las métricas precisión y sensibilidad fue el modelo de redes neuronales artificiales (96.5% y 88.2% para la clasificación de los niveles alto y medio), seguido del modelo de bosques aleatorios (96.8% para la clasificación del nivel bajo).

La calidad del aire si presenta una influencia en el desarrollo urbano sostenible. Tanto en la predicción con información temporal como espacial, los indicadores de calidad del aire: media anual de material particulado PM_{10} y $PM_{2.5}$ establecen un alto nivel de influencia en el proceso de clasificación; el PM_{10} ejerce la mayor influencia y el $PM_{2.5}$ un 80.5% en el modelo de redes neuronales. Llama la atención la agrupación de variables ambientales más influyentes en la clasificación, además del material particulado, se incluye el índice de vegetación normalizado (48.6%); se suman al análisis los efectos de la expansión urbana a través de la variación en vegetación en la comparación entre el año 2000 y 2016, con un nivel de importancia en la clasificación del 65.5%.

De la aplicación de las herramientas se concluye además que la elección apropiada de un modelo de clasificación debe reconocer el comportamiento de sus métricas de desempeño como un conjunto, identificando los aspectos que podrían sacrificarse al anteponer una métrica sobre otras; adicionalmente, dichos resultados deben contrastarse con el comportamiento de la información de entrada al modelo. La aplicación de métricas como la exactitud balanceada, la precisión, la especificidad y la sensibilidad permitió identificar las herramientas de mejor desempeño, incluida la aplicación de la media armónica entre las métricas precisión y sensibilidad.

Por su parte las características propias de planeación de los territorios urbanos, en cuanto a la ejecución de planes, programas y proyectos, no vinculan el reporte de información bajo una escala temporal diaria, sino que los reportes se realizan en algunos casos a una escala mensual y en otros anual. El análisis espacial, además de proveer información detallada y clara del territorio, resuelve las dificultades de acceso a grandes volúmenes de información para la calibración y entrenamiento de los modelos de aprendizaje automático. No obstante, para proveer un análisis del desempeño sostenible consistente, se hace necesario conjugar la información para cada periodo con los indicadores elegidos en la evaluación espacial. Se destaca como limitante que, en el procesamiento de la información espacial para los diferentes indicadores, es posible que se presente pérdida de información como resultado del proceso de rasterización; dicha situación puede tratarse mediante el cálculo de densidades con la aplicación del método IDW.

El desarrollo urbano sostenible es un reto de retos, incluye superar los desafíos en las dimensiones que le soportan, ambiental, social, económica, institucional y sus interacciones. Además, el crecimiento urbano supone diversos elementos que desafían la sostenibilidad, por lo que aún con la definición de objetivos y metas a nivel global y con el compromiso de las naciones, los problemas deben resolverse desde el nivel micro territorial, pues se presenta una imagen más cercana y real a cada realidad.

Finalmente, este trabajo concluye con tres mensajes centrales:

1. La clasificación de los niveles de avance en el desarrollo sostenible en el análisis espacial y temporal, a partir de herramientas de aprendizaje automático, permitió identificar las variables de mayor influencia en el proceso de clasificación; entre estas, los indicadores que informan del estado de la calidad del aire y otros que se correlacionan, como lo son los efectos de la expansión urbana.
2. La clasificación espacial de los niveles de sostenibilidad, desde el enfoque de micro territorios, no es homogénea; en algunas zonas confluyen características que califican el deterioro ambiental, en otras zonas confluyen mejores condiciones socioeconómicas y de habitabilidad de los espacios. En este orden de ideas, la caracterización de cada dimensión a nivel temporal y espacial permite identificar de manera diferenciada oportunidades de mejora, para avanzar en un mayor equilibrio de las interacciones habitable, viable, equitativa y las propias dimensiones del desarrollo sostenible.
3. En desarrollo de los objetivos específicos planteados para este trabajo doctoral se formuló una metodología cuya implementación secuencial permite identificar el grado de influencia de la calidad del aire, en el desarrollo urbano sostenible, analizada en virtud de sus indicadores más relevantes.

5.2 Desarrollos futuros

Estudios futuros podrían incluir un análisis de escenarios y brechas, identificando para cada indicador, según su variable de importancia, los elementos que podrían trabajarse y la variación que esas modificaciones podrían ejercer en el nivel de desempeño sostenible de los territorios. Se resalta además que, a pesar de que la disponibilidad de información es una continua limitación para el desarrollo de estudios de estas características, es recomendable continuar avanzando en el análisis de micro territorios, orientando acciones en el marco de la gobernanza, para establecer la implementación de estrategias, planes, programas y proyectos desde el nivel más elemental en el marco territorial.

Adicionalmente, la jerarquización de variables de mayor importancia, en el proceso de clasificación con las herramientas de aprendizaje automático, es un insumo necesario para la planificación, priorización de medidas y presupuestos desde la dimensión institucional.

Para afrontar el reto de la disponibilidad de datos en los micro territorios se requiere de información de acceso al público y de continuo reporte, que incluya el registro de datos para los diferentes indicadores asociados a la Agenda 2030. La continuidad en el registro de información espacio temporal, así como de la alimentación de observatorios de información con reportes más frecuentes que la escala anual contribuye en el monitoreo de los territorios y en la definición de acciones de mejora; la operación de redes de información ciudadana, con la participación de la academia y diversos actores puede contribuir con el registro, monitoreo y reporte de información de abajo hacia arriba, aportando en un mejor conocimiento de los territorios, a un menor costo. Algunos trabajos vinculan el uso de equipos y plataformas al alcance de la ciudadanía para informar sobre la calidad de los

recursos naturales, o de variables en la dimensión social y económica; estos productos podrían involucrarse en los observatorios y bases de información zonal, previa evaluación de calidad.

Otros desarrollos futuros incluyen:

- La aplicación de la metodología generada en este trabajo en diferentes escalas territoriales; un desarrollo futuro comprende la comparación de los resultados generados frente al comportamiento de otros micro territorios, ciudades y regiones, identificando además las variables con mayor influencia en el proceso de clasificación. La aplicación de la metodología diseñada como herramienta comparativa del desempeño de un territorio frente a otro, incluso la comparación de su propio desempeño en el tiempo proporcionará la información necesaria en la planificación no sólo en cada tema y dimensión del desarrollo sostenible, sino además desde una visión más global como insumo de planificación gubernamental.
- La habilitación de una plataforma en la que se incluya el pronóstico continuo del desarrollo urbano sostenible a partir de herramientas de aprendizaje automático; su operación requiere la articulación con la información publicada en plataformas estadísticas, observatorios de desarrollo sostenible, entre otros.
- El desarrollo e inclusión de un modelo de dispersión de contaminantes atmosféricos para el micro territorio, basado en información específica de fuentes de emisión y condiciones meteorológicas. Se requiere de un inventario de emisiones, así como el fortalecimiento en el monitoreo de la calidad del aire del micro territorio.
- Teniendo en cuenta que la calidad del aire presenta importante influencia en el desempeño sostenible de un territorio, podría realizarse un análisis involucrando la contaminación de fondo, al ser un factor crónico que afecta la salud de la población.
- Un análisis del ausentismo laboral asociado a enfermedades del sistema respiratorio en las zonas con mayor posibilidad de presentar este tipo de enfermedades, según el pronóstico generado con Random forest. Este análisis puede además incluirse en un análisis de desempeño económico del micro territorio.
- Un análisis desde el punto de vista de otras afecciones a la salud relacionadas con la contaminación del aire: enfermedades cardiopulmonares, mentales y/o dérmicas.

6 Capítulo 6 Referencias

Antanasijević, D., Pocajt, V., Ristić, M., Perić-Grujić, A., 2017. A differential multi-criteria analysis for the assessment of sustainability performance of European countries: Beyond country ranking. J. Clean. Prod. 165, 213–220. https://doi.org/10.1016/j.jclepro.2017.07.131

Banco Mundial, 2020. Desarrollo urbano [WWW Document]. URL https://www.bancomundial.org/es/topic/urbandevelopment/overview (accessed 10.23.20).

Boivin, M., Tanguay, G.A., 2018. How Urban Sustainable Development Can Improve Tourism Attractiveness. ARA J. Tour. Res. / Rev. Investig. Turística 8, 53.

Carrillo-Rodríguez, J., Toca, C.E., 2013. Desempeño sostenible en Bogotá: Construcción de un indicador a partir del desempeño local. Eure 39, 165–190. https://doi.org/10.4067/S0250-71612013000200008

Cui, X., Fang, C., Liu, H., Liu, X., 2019. Assessing sustainability of urbanization by a coordinated development index for an Urbanization-Resources-Environment complex system: A case study of Jing-Jin-Ji region, China. Ecol. Indic. 96, 383–391. https://doi.org/10.1016/j.ecolind.2018.09.009

DANE, 2020. Medida de pobreza multidimensional de fuente censal-información a nivel de manzana [WWW Document]. URL http://geoportal.dane.gov.co/visipm/

Dinov, I.D., 2018. Data science and predictive analytics: Biomedical and health applications using R, Data Science and Predictive Analytics: Biomedical and Health Applications using R. https://doi.org/10.1007/978-3-319-72347-1

Espejo, D.H., 2010. Noción y elementos de la justicia ambiental: directrices para su aplicación en la planificación territorial y en la evaluación ambiental estratégica. Rev. Derecho XXIII, 9–36.

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S.H., Rodríguez-Roda, I., Holmes, G., 2018. Environmental Modelling & Software Which method to use ? An assessment of data mining methods in Environmental Data Science. Environ. Model. Softw. 110, 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021

Gómez-Losada, Á., Santos, F.M., Gibert, K., Pires, J.C.M., 2019. Computers , Environment and Urban Systems A data science approach for spatiotemporal modelling of low and resident air pollution in Madrid (Spain): Implications for epidemiological studies. Comput. Environ. Urban Syst. 75, 1–11. https://doi.org/10.1016/j.compenvurbsys.2018.12.005

Gounaridis, D., Chorianopoulos, I., Koukoulas, S., 2018. Exploring prospective urban growth trends under different economic outlooks and land-use planning scenarios: The case of Athens. Appl. Geogr. 90, 134–144. https://doi.org/10.1016/j.apgeog.2017.12.001

Härdle, W.K., Simar, L., 2014. Applied Multivariate Statistical Analysis, Fourth. ed.

Springer Berlin Heidelberg, Berlin, Germany. https://doi.org/10.1007/978-3-662-45171-7

Ifaei, P., Karbassi, A., Lee, S., Yoo, C., 2017. A renewable energies-assisted sustainable development plan for Iran using techno-econo-socio-environmental multivariate analysis and big data. Energy Convers. Manag. 153, 257–277. https://doi.org/10.1016/j.enconman.2017.10.014

Jenks, G.F., 1967. The data model concept in statistical mapping., in: International Yearbook of Cartography. p. 7: 186:190.

Kubat, M., 2017. An Introduction to Machine Learning, Second. ed, An Introduction to Machine Learning. Springer International Publishing, FL,USA. https://doi.org/10.1007/978-3-319-63913-0

Kuhn, M., Wing, J., Steve Weston, Andre, Williams, Keefer, C., Engelhardt, A., Cooper, T., 2013. Package ‘ caret .’

Lary, D.J., Lary, T., Sattler, B., 2015. Using Machine Learning to Estimate Global PM $_{2.5}$ for Environmental Health Studies. Environ. Health Insights 9s1, EHI.S15664. https://doi.org/10.4137/EHI.S15664

Li, S., Batterman, S., Wasilevich, E., Elasaad, H., Wahl, R., Mukherjee, B., 2011. Asthma exacerbation and proximity of residence to major roads: A population-based matched case-control study among the pediatric Medicaid population in Detroit, Michigan. Environ. Heal. A Glob. Access Sci. Source 10. https://doi.org/10.1186/1476-069X-10-34

Li, Y., Wu, Y.X., Zeng, Z.X., Guo, L., 2006. Research on forecast model for sustainable development of Economy-Environment system based on PCA and SVM. Proc. 2006 Int. Conf. Mach. Learn. Cybern. 2006, 3590–3593. https://doi.org/10.1109/ICMLC.2006.258576

Meijering, J. V., Tobi, H., Kern, K., 2018. Defining and measuring urban sustainability in Europe: A Delphi study on identifying its most relevant components. Ecol. Indic. 90, 38–46. https://doi.org/10.1016/j.ecolind.2018.02.055

Mesnard, L. De, 2013. Computers & Geosciences Pollution models and inverse distance weighting : Some critical remarks. Comput. Geosci. 52, 459–469. https://doi.org/10.1016/j.cageo.2012.11.002

Meyer, D., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C., 2019. Package ‘e1071.’

Mirshojaeian, H.H., Kaneko, S., 2011. Dynamic sustainability assessment of countries at the macro level: A principal component analysis. Ecol. Indic. 11, 811–823. https://doi.org/10.1016/j.ecolind.2010.10.007

Molina-Gómez, N.I., Calderón-Rivera, D.S., Sierra-Parada, R., Díaz-Arévalo, J.L., López-Jiménez, P.A., 2021a. Analysis of incidence of air quality on human health: a case study on the relationship between pollutant concentrations and respiratory diseases in Kennedy, Bogotá. Int. J. Biometeorol. 65, 119–132. https://doi.org/10.1007/s00484-

020-01955-4

Molina-Gómez, N.I., Díaz-Arévalo, J.L., López-Jiménez, P.A., 2021b. Air quality and urban sustainable development: the application of machine learning tools. Int. J. Environ. Sci. Technol. 18, 1029–1046. https://doi.org/10.1007/s13762-020-02896-6

Molina-Gómez, N.I., Rodríguez-Rojas, K., Calderón-Rivera, D., Díaz-Arévalo, J.L., López-Jiménez, P.A., 2020. Using machine learning tools to classify sustainability levels in the development of urban ecosystems. Sustain. 12, 3326. https://doi.org/10.3390/SU12083326

Nilashi, M., Rupani, P.F., Rupani, M.M., Kamyab, H., Shao, W., Ahmadi, H., Rashid, T.A., Aljojo, N., 2019. Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. J. Clean. Prod. 240, 118162. https://doi.org/10.1016/j.jclepro.2019.118162

OMS, 2018. Nueve de cada diez personas de todo el mundo respiran aire contaminado Sin embargo, cada vez hay más países que toman medidas [WWW Document]. URL http://www.who.int/es/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action (accessed 10.23.18).

Organización de las Naciones Unidas, 2018. Marco de indicadores mundiales para los Objetivos de Desarrollo Sostenible y metas de la Agenda 2030 para el Desarrollo Sostenible.

Pérez-Ortíz, M., de La Paz-Marín, M., Gutiérrez, P.A., Hervás-Martínez, C., 2014. Classification of EU countries' progress towards sustainable development based on ordinal regression techniques. Knowledge-Based Syst. 66, 178–189. https://doi.org/10.1016/j.knosys.2014.04.041

Prescott-allen, R., 1997. Barómetro de la Sostenibilidad.

Quiroga Martínez, R., Stockins, P., Holloway, M., Taboulchanas, K., Sanchez, A., 2009. Guía metodológica para desarrollar indicadores ambientales y de desarrollo sostenible en países de América Latina y el Caribe. CEPAL. United Nations. Economic Commission for Latin America and the Caribbean., Santiago de Chile.

Rajaonson, J., Tanguay, G.A., 2017. A sensitivity analysis to methodological variation in indicator-based urban sustainability assessment: a Quebec case study. Ecol. Indic. 83, 122–131. https://doi.org/10.1016/j.ecolind.2017.07.050

Ripley, B., Venables, W., 2016. Feed-Forward Neural Networks and Multinomial Log-Linear Models. February.

Rokach, L., Maimon, O., 2015. Data mining with decision trees : theory and applications, 2nd ed. World Scientific Publishing Co. Pte. Ltd. 5, Singapore.

Saaty, R.W., 1987. The analytic hierarchy process-what it is and how it is used. Math. Model. 9, 161–176. https://doi.org/10.1016/0270-0255(87)90473-8

Salam, M.T., Islam Talat, Gilliland, F.D., 2008. Recent evidence for adverse effects of residential proximity to traffic sources on asthma. Curr. Opin. Pulm. Med. 14, 3–8.

https://doi.org/10.1097/MCP.0b013e3282f1987a

Scipioni, A., Mazzi, A., Mason, M., Manzardo, A., 2009. The Dashboard of Sustainability to measure the local urban sustainable development : The case study of Padua Municipality. Ecol. Indic. 9, 364–380. https://doi.org/10.1016/j.ecolind.2008.05.002

Secretaría Distrital de Planeación, 2020. Secretaría Distrital de Planeación [WWW Document]. URL http://www.sdp.gov.co/transparencia/informacion-interes/glosario/unidades-de-planeamiento-zonal-upz (accessed 11.22.20).

Shen, L., Kyllo, J., Guo, X., 2013. An Integrated Model Based on a Hierarchical Indices System for Monitoring and Evaluating Urban Sustainability. Sustainability 5, 524–559. https://doi.org/10.3390/su5020524

Singh, R.K., Murty, H.R., Gupta, S.K., Dikshit, A.K., 2012. An overview of sustainability assessment methodologies. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2011.01.007

Singh, S.R., Murthy, H.A., Gonsalves, T.A., 2010. Feature Selection for Text Classification Based on Gini Coefficient of Inequality, in: Liu, H., Motoda, H., Setiono, R., Zheng, Z. (Eds.), The Fourth Workshop on Feature Selection in Data Mining. pp. 76–85.

Tajudin, M.A.B.A., Khan, M.F., Mahiyuddin, W.R.W., Hod, R., Latif, M.T., Hamid, A.H., Rahman, S.A., Sahani, M., 2019. Risk of concentrations of major air pollutants on the prevalence of cardiovascular and respiratory diseases in urbanized area of Kuala Lumpur, Malaysia. Ecotoxicol. Environ. Saf. 171, 290–300. https://doi.org/10.1016/J.ECOENV.2018.12.057

Torres-Delgado, A., López Palomeque, F., 2018. The ISOST index: A tool for studying sustainable tourism. J. Destin. Mark. Manag. 8, 281–289. https://doi.org/10.1016/j.jdmm.2017.05.005

Toumi, O., Le Gallo, J., Ben Rejeb, J., 2017. Assessment of Latin American sustainability. Renew. Sustain. Energy Rev. 78, 878–885. https://doi.org/10.1016/j.rser.2017.05.013

UN Commission on Sustainable Development, 2001. Indicators of sustainable development: Guidelines and Methodologies.

Wang, X., Xiao, Z., 2017. Regional eco-efficiency prediction with Support Vector Spatial Dynamic MIDAS. J. Clean. Prod. 161, 165–177. https://doi.org/10.1016/j.jclepro.2017.05.077

Whitehead, M., 2003. (Re)analysing the sustainable city: Nature, urbanisation and the regulation of socio-environmental relations in the UK. Urban Stud. 40, 1183–1206. https://doi.org/10.1080/0042098032000084550

Zeng, L., Guo, J., Wang, B., Lv, J., Wang, Q., 2019. Analyzing sustainability of Chinese coal cities using a decision tree modeling approach. Resour. Policy 64, 101501. https://doi.org/10.1016/j.resourpol.2019.101501

Zhang, Y., Huan, Q., 2006. Research on the evaluation of sustainable development in Cangzhou city based on neural-network-AHP, in: Proceedings of the Fifth International Conference on Machine Learning and Cybernetics. pp. 3144–3147.

https://doi.org/10.1109/ICMLC.2006.258407

Zhang, Y., Shang, W., Wu, Y., 2009. Research on sustainable development based on neural network, in: 2009 Chinese Control and Decision Conference. IEEE, pp. 3273–3276. https://doi.org/10.1109/CCDC.2009.5192476

Apéndices

Apéndice A.

Analysis of incidence of air quality on human health. A case study on the relationship between pollutant concentrations and respiratory diseases in Kennedy, Bogotá

Documento versión del autor publicado en la Revista: International Journal of Biometeorology (Electronic ISSN 1432-1254; Print ISSN 0020-7128). Recibido el 13 de Mayo de 2019; revisado el 10 de Junio de 2020; aceptado el 12 de Junio de 2020; publicado el 13 de Julio de 2020. DOI: https://doi.org/10.1007/s00484-020-01955-4

Nidia Isabel Molina-Gómez[1,2], Dayam Soret Calderón-Rivera[1], Ronal Sierra-Parada[1], José Luis Díaz-Arévalo[3], P. Amparo López-Jiménez[2]

[1.]Department of Environmental Engineering, Universidad Santo Tomás, Bogotá, Colombia

[2.]Hydraulic and Environmental Engineering Department, Universitat Politècnica de València, Valencia, Spain

[3.]Department of Civil and Agricultural Engineering, Universidad Nacional de Colombia, Bogotá, Colombia

---

Mediante el uso de aprendizaje automático, información espacial de contaminantes atmosféricos, variables meteorológicas y condiciones sociodemográficas fue posible identificar las variables más relevantes y las zonas de interés frente a la posibilidad de presentar una enfermedad del sistema respiratorio en el microterritorio.

ABSTRACT

Thousands of deaths associated with air pollution each year could be prevented by forecasting the behavior of factors that pose risks to people's health and their geographical distribution. Proximity to pollution sources, degree of urbanization, and population density are some of the factors whose spatial distribution enables the identification of possible influence on the presence of respiratory diseases (RD). Currently, Bogotá is among the cities with the poorest air quality in Latin America. Specifically, the locality of Kennedy is one of the zones in the city with the highest recorded concentration levels of local pollutants over the last 10 years. From 2009 – 2016, there were 8619 deaths associated with respiratory and cardiovascular diseases in the locality. Given these characteristics, this study set out to identify and analyze the areas in which the primary socio-economic and environmental conditions contribute to the presence of symptoms associated with RD. To this end, information collected in field by performing georeferenced surveys was analyzed through geostatistical and machine learning tools which carried out cluster and pattern analyses. Random forests and AdaBoost were applied to establish hotspots where RD could occur, given the conjugation of predictor variables in the micro-territory. It was found that random forests outperformed AdaBoost with 0.63 AUC. In particular, this study's approach applies to densely populated municipalities with high levels of air pollution. In using these tools, Municipalities can anticipate environmental health situations and reduce the cost of respiratory disease treatments.

Keywords: geostatistics, machine learning, sustainable development, air quality, hot spots

INTRODUCTION

In developing countries, air pollution is among the environmental problems of greatest concern. It is a risk factor for populations' health, which can affect different age groups more severely. Furthermore, growing urbanization has increased urban density and populations' proximity to pollution sources. Therefore, it is essential to analyze the impact of atmospheric pollutants on a population's health. In this vein, epidemiological studies and predictive modeling which employ machine learning (ML) techniques have been carried out.

Epidemiological studies consist of designing experimental or observational studies. Experimental studies are randomized and quasi-experimental trials, in which the researcher has a certain degree of control over the variables. Observational studies include cohort, case-control, cross-sectional and ecological studies (Kestenbaum, 2019). In cohort studies, individuals are classified in sub-groups, according to exposure to a potential cause of sickness, in which the entire evolution of the cohort is monitored (Lazcano-Ponce et al., 2000). In case-control studies, a comparison is made between the groups in which the event occurs, and those in which it does not. Cross-sectional studies analyze the frequency of a health event with respect to the exposure level of the analyzed individuals or group in a given moment (Hernández et al., 2000). Ecological, correlational and exploratory, studies focus on studying groups with an analysis of geographic areas or different time periods, and are useful in evaluating multiple exposure levels (Borja-Aburto, 2000).

ML techniques were also used to identify the influence of physical and chemical factors in the population's health. ML can process large volumes of data, as well as linear and non-linear relationships (Ivanov 2018). ML can also perform classification and regression tasks through decision trees, artificial neural networks (ANN), support vector machines (SVM) or through ensemble methods, such as random forests (RF) or adaptive boosting (AdaBoost-AdB). From a data set, ML is able to identify data patterns and predict their behavior (Kuhn and Johnson 2013). ANN were applied to determine the influence of physical and chemical stressors in hospital admissions for respiratory and cardiac diseases (Kassomenos et al., 2011; Polezer et al., 2018). Generalized boosting models were applied in the same manner for exposure periods before, during and after forest fires (Reid et al., 2016). Furthermore, Bayesian kernel regression was used to estimate the function of response doses and to identify the combination of pollutants responsible for adverse health effects (Bobb et al., 2015). Moreover, statistical tools such as generalized linear regression, multiple linear regression, logistic regression and ML techniques (RF, SVM, and ANN) were used to forecast atmospheric pollutant levels that may generate a public health risk (Huang et al., 2018; Ivanov et al., 2018; Kami, 2019; Ni et al., 2017; Pandey et al., 2013; Weizhen et al., 2014; Zhan et al., 2017).

The above referenced studies forecasted pollutants' behavior and health effects from information recorded in a database. However, there is still a lack of forecasting of possible respiratory disease (RD) hotspots based on variables' spatial distribution and behavior. The spatial distribution of risk factors and their interactions in territories increase interest in knowing future spatial scenarios of possible health effects. The use of ML and spatial zoning of these factors facilitate forecasting variables' behavior by identifying hotspots with a territorial approach, which is more specific than national or capital city focuses. These scenarios are essential for decision-makers so that they are able to implement measures to mitigate costs related to the treatment of morbidity and mortality.

With nearly 8.3 million inhabitants and located along the plateau of the eastern range of the Colombian Andes at an elevation of 2600 meters above sea level (m.a.s.l), Bogotá is one of the most populated cities in Latin America and one of the cities with the highest recorded levels of atmospheric pollutants, which represent a risk factor for its population. 21.5% of medical consultations performed for the productive age population (15 – 65 years old) are related to air pollution (García-Ubaque et al., 2011). Furthermore, changes in NO_2, SO_2 and $PM_{2.5}$ concentration levels in Bogotá were correlated with statistically significant effects regarding changes in emergency room visits due to RD by children younger than fifteen years old, while changes in SO_2, PM_{10} and $PM_{2.5}$ were related to changes in emergency room visits for circulatory system diseases in adults older than sixty years (Rodríguez-Villamizar et al., 2018).

In addition to the above, several studies have been developed in Bogotá aimed at forecasting its air quality. A combined linear regression model was used to predict air quality (Westerlund et al., 2014). Moreover, ANN were applied to predict PM_{10} and $PM_{2.5}$ concentration levels (Franceschi et al., 2018). As a result of the data

analyzed, it was determined that Kennedy is one of the zones in the city with the highest air pollution levels, which also happens to be one of the most densely populated localities in Bogotá.

Among the references consulted, there was no study that combined geostatistical tools and ML to identify specific zones in which cases of RD may occur due to air quality. Therefore, an ecological case study was conducted by applying these tools in a specific analysis of Kennedy, setting out to determine the influence of meteorological variables and atmospheric pollutants on the population's health, establishing not only the most relevant variables, but also the zones of greatest interest in geographic spaces based on the locality's characteristics. Applying geostatistical tools and ML in an environmental health study on an atmospheric component is innovative. Furthermore, the local scale of the analysis is emphasized, in addition to data collection in the field being one of the inputs to feed the ML model. This is the first study of this nature developed in one of the most populated zones of a city such as Bogotá. The approach created in this study can be applied to different territories, particularly densely populated areas with high air pollution levels. Different municipalities can anticipate environmental health situations and reduce the cost of RD treatments by applying these tools.

MATERIALS AND METHODS

This study consists of three sequentially phases which are described below (see Fig. 1), namely: study area analysis, exploratory analysis of air quality and RD, and the forecasting model with ML techniques and geographical information system (GIS).

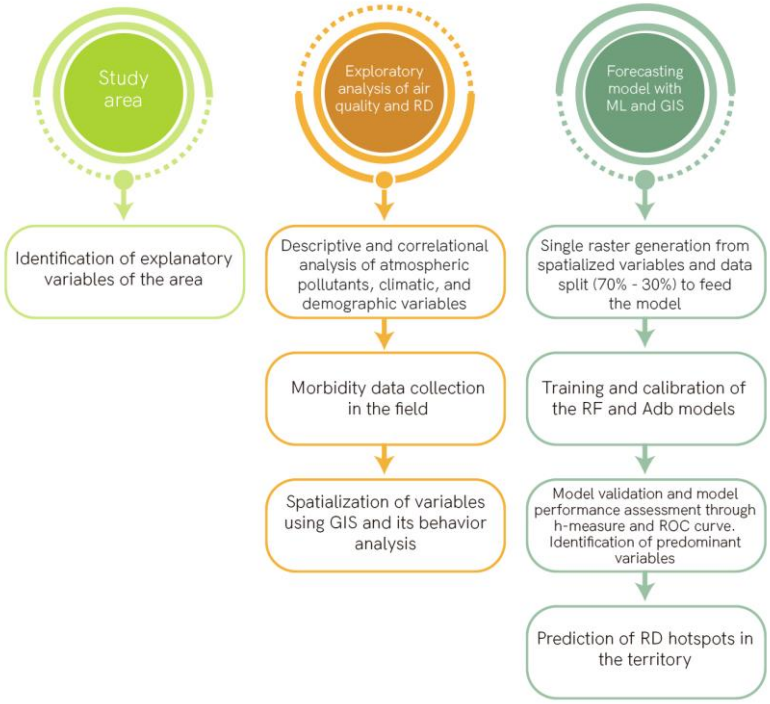


Fig. 1 Methodological framework to identify and forecast RD hotspots.

STUDY AREA

The study area is Kennedy, located in southwest Bogotá, Colombia. It covers an area of 38.6 km^2, of which, 1.9% is green space and 10% is protected. Kennedy is located in a transition zone between the eastern plateau and mountains. It is a flat zone which borders five Bogota localities (Puente Aranda, Fontibon, Bosa, Ciudad Bolívar and Tunjuelito), and is where some of the city's main industrial, mining, and commercial activities are located. It also borders Mosquera, Cundinamarca. Wind in Kennedy predominately comes from southwest

Bogotá at speeds of 2.2 to 2.5 m/s. The average temperature is 14.6 ± 0.4 ºC, with its highest recorded values in 2016 (14.9 ± 0.8 ºC). Moderate thermal inversions are common, primarily in the dry months. With respect to precipitation, low values have been recorded in this area of the city, with cumulative averages for the period 2009 – 2016 between 483 and 1018 mm, with a multi-annual average precipitation of nearly 767 mm (Distrital Secretariat of the Environment, 2017).

This locality is made up of twelve zoning planning units (ZPU), which act as territorial units for urban development planning at the zonal level: Américas, Bavaria, Calandaima, Carvajal, Corabastos, Castilla, Gran Britalia, Kennedy Central, Las Margaritas, Patio Bonito, Tintal Norte and Timiza. Four of the above are for residential urban land use,[1] three are for residential use in incomplete urbanization zones,[2] two are in the developing stages,[3] one is the urban center of the locality,[4] and two are allocated for public use[5] (SDP, 2018). There are units registered for industrial development activities in the Américas, Carvajal and Bavaria ZPUs (Galindo, 2013), and approximately 47.2% of households are located near industries (DANE, 2018).

Roadways with high vehicular traffic cross the locality, such as Avenida Boyacá, Avenida Ciudad de Cali, and Avenida de las Américas (see Fig. 2), on which light, cargo and public transportation vehicles represent the main traffic. Predominantly type 1 and 2 roadways cross and border the locality, as part of the arterial road system. These roads support traffic flows caused by the inter-urban transport of goods and people. Due to their length and characteristics they support traffic caused by mass public transportation and connect to the local road network. Type 1 roads are 60 m wide, while type 2 are 40 m wide. Furthermore, there are local roads within the study area with widths ranging from 4 to 22 m that facilitate entry and local traffic caused mainly by individual transport vehicles. The locality is also bordered by the Autopista Sur highway and Calle 13, whose road accesses to the city were used by an average of 5300 – 6600 trucks in 2017, with a daily average of 12,000 different types of vehicles going to the CORABASTOS supply center, according to government entities in Bogotá.

The predominant buildings in the locality are made of concrete, cinder blocks and bricks, whose heights mainly range from 1 to 5 floors; 67% correspond to buildings up to 3 floors (8.1 m) high, with 18.5% of buildings having 4 and 5 floors (up to 13.5 m), and some buildings are taller than 37 m (DANE, 2018).

Kennedy has an estimated 1,208,980 inhabitants according to the population census (SDP, 2018). In accordance with a 2017 analysis of its population structure, 53.7% of its population are adults, while the early childhood and adolescent population groups have a smaller representation. Furthermore, the overall population rate in the labor market in Bogotá is approximately 60.8%[6] (SDP, 2018).

[1] The use changes are occurring in predominately residential sectors with an increase of unplanned territorial occupancy.
[2] Strata 1 & 2 non-consolidated peripheral residential sectors with deficiencies in their infrastructure, accessibility, equipment, and public space.
[3] Under-developed sectors with large unoccupied lots.
[4] Consolidated sectors that have urban centers with the dominating residential use having been displaced for uses that encourage economic activities.
[5] Large areas allocated to produce urban and metropolitan equipment.
[6] The working age population is 12 years and older in the urban zone, which for Kennedy corresponds to 1,019,894 people.

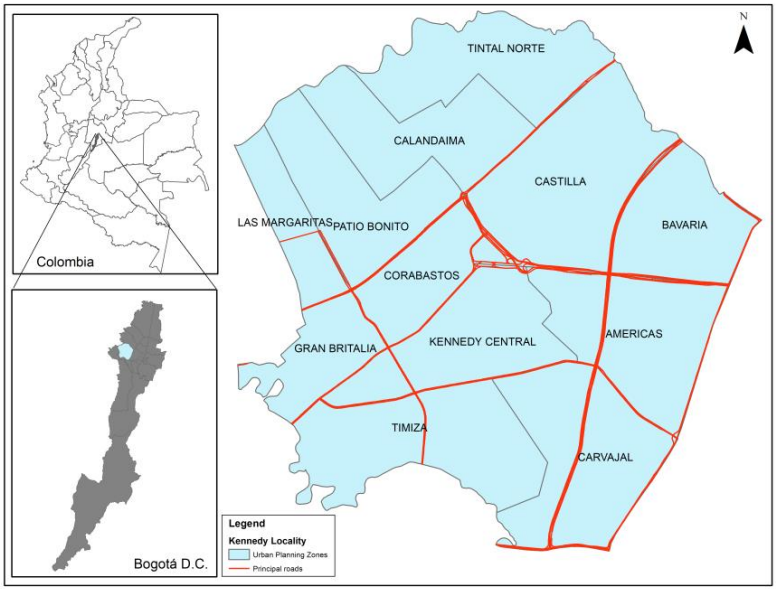


Fig. 2. Locality of Kennedy and its location in Bogotá, Colombia. The twelve ZPUs that make up the internal distribution of Kennedy along with its main roadways are displayed.

Air quality and its effects on health

Bogotá has an Air Quality Monitoring Network (AQMN), which is comprised of thirteen monitoring stations in the city's urban area, two of which are located in Kennedy (see Fig. 3). The first (Kennedy station) is situated in a residential zone and monitors PM_{10}, $PM_{2.5}$, NO, NO_2, CO, SO_2, as well as meteorological variables including humidity, barometric pressure, solar radiation, temperature, precipitation, wind speed and direction. The second (Carvajal station) is located in a residential zone with a presence of industrial activity. It is an industrial-traffic station that monitors PM_{10}, $PM_{2.5}$, NO, NO_2, NOx, O_3, CO, SO_2 and meteorological variables such as precipitation, temperature, wind direction and speed.

This study analyzed data from the two stations mentioned above, as well as the Tunal, Puente Aranda and Centro de Alto Rendimiento stations (see Fig. 3), located in Kennedy's influence area, which are the traffic, industrial and background stations, respectively. These automated stations monitor the same parameters mentioned above as the stations in Kennedy. Furthermore, the Mosquera-Sena manual station is located in the Bogotá Savanna, and measures PM_{10}, SO_2 and NO_2. The Centro de Alto Rendimiento station is located in a zone with a low concentration of pollutants, where winds from all directions converge, and has historically recorded low pollution levels. The Tunal and Puente Aranda stations are located in zones with high traffic and industrial activities. The Tunal station receives winds from the south, while the winds that hit the Puente Aranda station come from the west and northwest (Distrital Secretariat of the Environment, 2017).

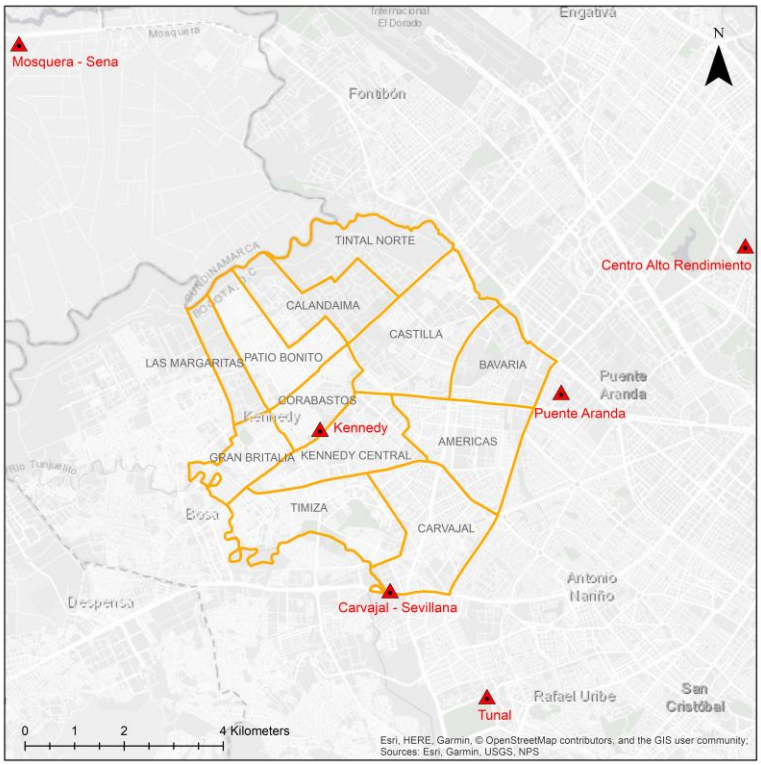


Fig. 3. Location of the monitoring stations in the study site and the area of influence.

In 2015, the Carvajal and Kennedy stations had the highest concentrations of PM_{10} and $PM_{2.5}$ in the city. In the first quarter of 2019, environmental emergencies were declared in areas monitored by these two stations. These emergencies occurred as the result of a variation in meteorological conditions and the intensification of the temperature inversion phenomena during the dry season, which generally occurs in the city with moderate effects and breaks atmospheric stability between 7:00 – 9:00 a.m. From 2011 – 2015, neither station met the national regulation standard and recorded yearly concentration averages that were among the highest for stations that monitor PM_{10} and $PM_{2.5}$ in the country, which met the temporal coverage criterion of 75% (IDEAM, 2016). NO_2, O_3, CO and SO_2 pollutants did not exceed the limits established in the regulation. However, SO_2 did have higher concentrations in the monitoring stations situated in the locality. The presence of different types of industrial activities in the city and in neighboring municipalities, as well as road and traffic conditions with different types of vehicles, all contribute to the increased concentration of atmospheric pollutants in Kennedy. It is important to note that in Bogotá, prevailing winds from the northeast and southeast displace particulate matter towards the west (Ramírez et al., 2018).

EXPLORATORY ANALYSIS

Work began on an exploratory analysis of the variables, which in terms of air quality, could impact the population's health in Kennedy. A descriptive analysis was conducted of the spatial distribution of atmospheric pollutants and meteorological variables for the period 2009 – 2017, as well as descriptive analyses of individual records from health care providers (RIPS, as per its Spanish acronym) in Kennedy, reported by the District Health Secretariat (DHS) for the same period, for diseases associated with air quality, in accordance with Version 10 of the International Classification of Diseases (ICD). Furthermore, a bivariate Pearson correlation of the continuous variables (pollutants, climatology, and demographic variables) was conducted for 2016. The data on atmospheric pollutants and meteorological variables was determined based on a weighted average calculated by the ArcGis 10.5.1 software, using information from the AQMN stations (Carvajal, Kennedy, Puente Aranda, Centro de Alto Rendimiento and Tunal), as well as the Mosquera-Sena station in Cundinamarca, which are in the locality and its boarding zones (see Fig. 3).

The spatial distribution of PM_{10}, $PM_{2.5}$, CO, NO_X, and SO_2 pollutants, as well as the precipitation and temperature meteorological variables were analyzed via the deterministic method for interpolation called inverse distance weighting (IDW) interpolation. This is a univariate interpolation method, which is useful in evaluating small study areas. To generate a predictive surface, the value taken by an unknown point is influenced more by nearby sampled data, than by data from areas further away (Ly et al. 2011). This method does not consider spatial groupings and has better results when the sampled data comes from irregularly spaced locations (Li and Heap 2014). This is the case of Bogotá, which has approximately one station every 23 km^2. Given that the information for this study comes from irregularly distributed monitoring points (see Fig. 3), this study contemplated examining the influence of the data recorded at the stations concerning its surrounding areas. The spatial behavioral analysis of the data was performed via the natural grouping data classification method proposed by Jenks (Jenks 1967).

FORECASTING MODEL WITH ML AND GEOGRAPHICAL INFORMATION SYSTEM (GIS)

Medical consultation records from RIPS do not contain information on the spatial location of health care service users. Consequently, using data provided by the DHS to spatially identify the zones of the locality with possible RD due to the presence of atmospheric pollutants was not possible. As such, the decision was made to develop the field work by conducting a survey on health perception, identifying individuals from households in the locality who have been diagnosed with a RD[7] in 2016. To this end, a georeferenced primary source data collection instrument was applied, which considered socio-demographic variables and the surveyed person's perception of their health condition. Through a structured questionnaire, the survey developed which consisted of thirty-one questions, was conducted with households in the locality. This instrument was applied in twelve ZPUs in 2017 in accordance with the sample size established by the study.

The required sample was established to conduct the surveys based on the 2016 number of inhabitants in each ZPU (SDP, 2018). The equation for finite populations was used with the following criteria: error: 4%; confidence level: 96%; and positive and negative variables: 50%. This equation was applied to the general population of the locality yielding a result of 656, which was distributed in accordance with the population proportion of each ZPU. During the development of the study, the sample size grew to 912 surveys, thus expanding its spatial coverage (see Fig. 4).

[7] According to the ICD, RD range from rhinopharyngitis, known as the common cold, to respiratory disorders in diseases classified elsewhere (J00-J99).

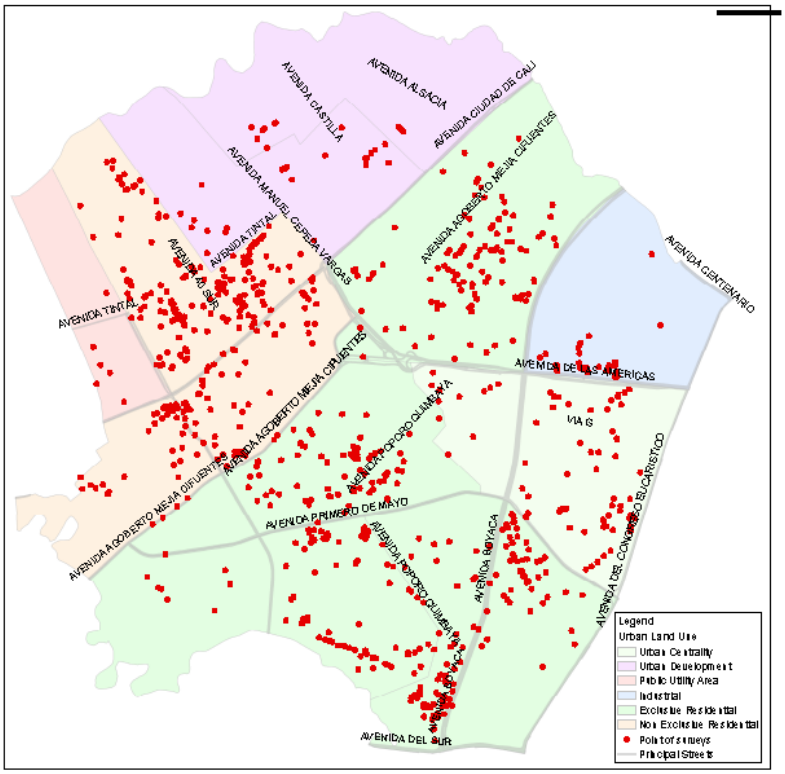


Fig. 4. Field data collection points in the Kennedy ZPUs, primary roadways, and land use typologies

Design and development of the forecasting model

Health risks arise from a combination of socio-economic factors, environmental conditions, habitat, and individual behavior. Geospatial and ML tools were applied to identify the areas of greatest interest related to the population's respiratory health, in contrast with the presence of pollution sources, pollutant distribution, and the exposed population. The ArcGis 10.5.1 software and the open-source software R 3.5.2 (CRAN 2018) were used for this purpose. Information was entered on atmospheric pollutants (PM_{10}, $PM_{2.5}$, CO, NOx, SO_2), meteorological variables, precipitation and temperature previously determined by the IDW method for 2016, as well as data on population, population density, households' proximity to roadways (type 1:T1 and type 2:T2), and land use typology. According to Salam et al. (2008) and Li et al. (2011), the proximity of households, located between 100 m – 1000 m to local and main roads, may increase the risk of RD. This study considered households' distance to primary and secondary roads (T1 and T2, respectively).

Geocoded information was also entered and arranged in a GIS of the set of categorical responses from the individuals surveyed to the question, "In the last year (2016), have you or any of members of your household been diagnosed by a doctor with a respiratory disease or infection such as asthma, pneumonia or severe lung disease?" A single raster was created with the resulting information, in which the analyzed variables (continuous and categorical) were overlaid and then used as input information for the R software. This process provided information on the twelve explanatory variables and the categorical answer for each point in the locality. RF and the AdB algorithm were the ML tools used, both of which improve the accuracy of single decision tree classifiers by combining trees grown (Breiman 2001). These tools maintain a bias-variance trade off trough bagging or boosting methods. It is important to note that ANN and SVM are also useful in classification tasks. However, collinearity of variables is a condition that limits the accuracy and generalization capacity of ANN (Kuhn and Johnson 2016). Furthermore, the proximity between classes in the geographical space limits the accuracy of SVM. RF and AdB perform better in those aspects and facilitate the analysis of variables distributed in space, which is useful in the integrated and spatial analysis of possible health risk factors.

RF are one of the most accurate bagging methods. RF are a consistent classifier in collecting tree-structured classifiers {h(x,(Ξ)k), k = 1,...}, in which {(Ξ)k} are independent random vectors identically distributed, with each tree issuing a single vote for the most popular class in the x input (Breiman, 2001). For categorical predictions, the voting process selects the class with the most votes (Kuhn and Johnson 2016). RF can handle large numbers of features (Ivanov et al. 2018) and identify the most important variables for the model. The precision of RF depends on the strength of the individual classifiers and the dependence measure between them (Breiman, 2001).

The model used 70% of the data for training and 30% for testing. The partition was performed by randomly considering the proportionality between affirmative and negative responses. A forecast was created of the areas with the strongest confluence of affirmative responses to the possibility of RD cases by a majority vote, resulting in the classification that determined the most influential variables in the model and the distribution of response data according to the conjugate of the predictor variables in the classification with RF. The model calibration included an iteration of 300 – 1500 trees, with every 100 trees establishing the best combination with the number of variables, according to the accuracy results and the Kappa index.

Subsequently, the AdB algorithm, which has no random elements and uses decision trees as the model base, was applied to create a strong classifier (an ensemble of trees) built from weak classifiers by successively reweighing them (Breiman, 2001). AdB is one of the most widely-used boosting methods in which each classifier focuses on the data that was erroneously classified by its predecessor, in order to adapt the algorithm and generate better results with each iteration and reduce the generalization error (Schapire and Freund, 2012)(Breiman, 2001). In this method, each constructed tree depends on its predecessor's trees and the prediction come from the most frequent selected class. The samples that are incorrectly classified in the iteration are given more weight than the samples correctly classified. Therefore, samples that are difficult to classify are given greater weight until Adb identifies the best model (Kuhn and Johnson 2016). In this study, the same input parameters for the RF model were used.

By using the *Mean Decrease Accuracy* tool, the variables with the greatest influence on the classification error were determined for each model. Subsequently, forecasts were made with 100% of the spatial behavior data according to possible RD cases in the locality. The H-measure and the classification error from the receiver operating characteristic (ROC) curve were used as the performance indicators. The H-measure is a measurement of the loss from erroneous classification contingent on the relative proportion of the objects belonging to each class (Hand, 2009). The ROC curve enables a comparison of the accuracy and precision of the representing model for each threshold value. This curve is a plot showing all the sensitivity and specificity pairs resulting from the continuous variation of cutoff over the entire range of observed results (Altman and Bland 1994). Furthermore, as a function of sensitivity and specificity metrics, the area under the ROC curve (AUC) is insensitive to disparities in class proportions. A perfect model separates the two classes with sensitivity and specificity values of 100% (Kuhn and Johnson 2016). Therefore, sensitivity and the specificity metrics of the diagnostic test, as well as the AUC closest to 1, in the 0.5 – 1.0 interval, represents greater accuracy than the discriminant test (Valle Benavides, 2017). This area establishes the probability that a random person with the disease has a higher measurement value than a random person without the disease (Altman and Bland 1994).

The "Geographic data analysis and modeling" raster and the "Bindings for the 'Geospatial' Data Abstraction Library" rgdal were the packages used in the R software to read and process the raster images. "Breiman and Cutler's random forests for classification and regression" were used to design and develop the RF model. The "C_lassification _A_nd _RE_gression _T_raining" caret was used to determine the most optimal model parameters. The "Visualizing the performance of scoring classifiers" ROCR was used to calculate the AUC and display the ROC curve. These packages made it possible to adjust the spatial information to the databases adapted for statistical and predictive processing. A computer with Core i5 8th generation processor, 8Gb RAM and 1Tb hard disk was used.

RESULTS

The locality of Kennedy is characterized by its location between primary roadways and the diversity of economic activities carried out in the same. It has gone through different changes as it is one of the most densely

populated localities in Bogotá. By using the IDW method for the period 2009 – 2017, a decreasing trend was found in the concentration of different pollutants with values ranging from: (78.72 – 53.11 ug/m^3) for PM_{10}; (35.09 – 24.32 ug/m^3) for $PM_{2.5}$; (1.2 – 0.73 ug/m^3) for CO; (64.15 – 40.46 ppm) for NO_x; and (8.69 – 2.77 ppb) for SO_2. The largest values mainly occurred in 2009, which also had the largest number of consultations associated with RD.

The PM_{10} and $PM_{2.5}$ values surpassed those established by WHO guidelines (PM_{10}= 20 µg/m^3; $PM_{2.5}$= 10 µg/m^3). According to the WHO (2006), these are the lowest levels that demonstrate, with more than 95% confidence, that total cardiopulmonary and lung cancer mortality increases in response to prolonged exposure to $PM_{2.5}$. However, values were recorded in 2017 that were close to WHO guideline values according to which, the risk of premature mortality is reduced by 6% compared to the severe level; (PM_{10} =50 µg/m^3; $PM_{2.5}$= 25 µg/m^3) (WHO, 2006). The SO_2, CO and NOx values indicate a reduction of pollutants. SO_2 did not surpass standards (30 ppb) set by the Environmental Protection Agency (EPA). In the case of CO, there is no yearly standard, yet the values recorded at the monitoring stations did not, at any time, exceed the Colombian standard (5000 µg/m^3), nor the EPA standard (9 ppm) for 8 hours of exposure. NO_X, which is an unregulated pollutant and ozone precursor, decreased by approximately 37% compared to the analyzed periods.

In the period covering 2009 – 2017, after the common cold, chronic obstructive pulmonary disease (COPD), acute bronchitis, and unspecified asthma were the most common RDs for which different patients went to consultations. Consultations ranged from: (1493 – 5744) for COPD; (2294 – 4736) for acute bronchitis; and (1380 – 2573) for unspecified asthma.

Correlation Analysis

A matrix was created by applying Pearson's method, which demonstrates high correlation between the 2016 climatic variables and atmospheric pollutants analyzed; precipitation and temperature have an inverse relationship (see Fig. 5). Moreover, in Fig. 5, the relationships with no significance are marked with an X, that is, their parameter *p-value* is greater than 0.05.

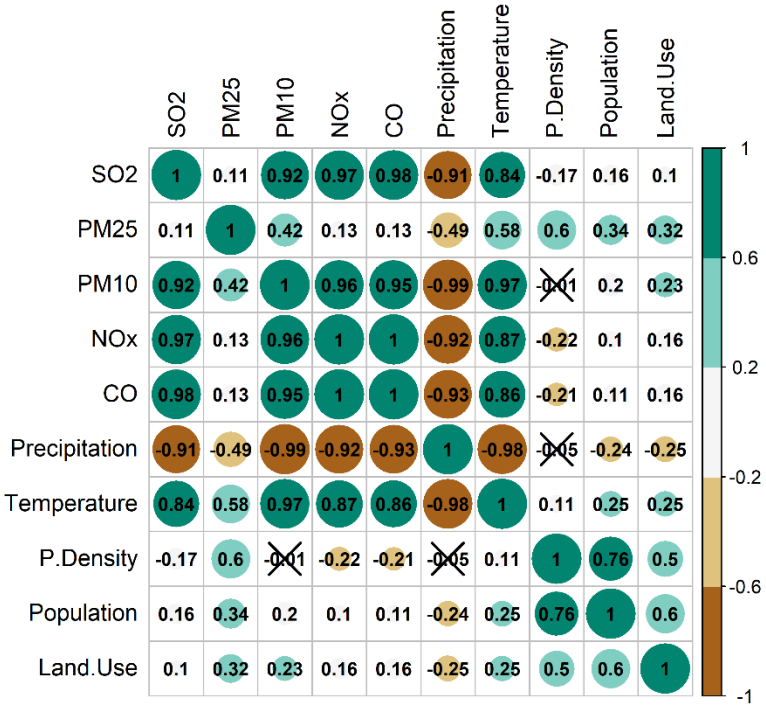


Fig. 5. Correlation matrix of pollutants, meteorological and demographic variables in 2016. The color scale on the sidebar shows the degree of positive (0 – 1) or negative (0 – -1) correlation between the variables.

Areas of interest for possible cases of respiratory disease

In 2016, the largest concentration of pollutants analyzed were found in the Carvajal and Timiza ZPUs (see Fig. 6). However, the highest concentrations of $PM_{2.5}$ were in the areas of the Corabastos, Kennedy Central, Carvajal, Patio Bonito, Calandaima, Margaritas, Gran Britalia and Timiza ZPUs in the center and western zones of the locality. Temperature had a constant behavior (14.98°C), with its lowest values found in the eastern zone of the Bavaria ZPU (14.7°C), which has larger precipitation values (769.3 mm) with respect to the rest of the study area (712.42 mm on average). The smallest precipitation values were found in the Carvajal and Timiza zones, with 620.6 mm and 637.2 mm, respectively.

In total, 912 surveys were conducted in the twelve ZPUs that make up the locality. The Patio Bonito, Carvajal, Kennedy Central and Castilla ZPUs had the largest number of affirmative responses to the questions asked in the field work (see Fig. 6); 21.4% of the individuals surveyed indicated that a member of their household was diagnosed with a RD, of which 51.8% corresponded to the working age population to 60 years old, 23.6% were young people between 5 – 14 years old, and 14.3% were people older than 60. Furthermore, it was found that 49.2% of respondents had lived in the study area for more than ten years. These indicative figures are comparable with those reported by DHS in 2016, given that in Kennedy nearly 27% of RD cases in children under 14 years of age were attended to in emergency rooms, and a prevalence of wheezing was reported in 12.6% of adults over 60 years old (District Health Secretariat, 2019).

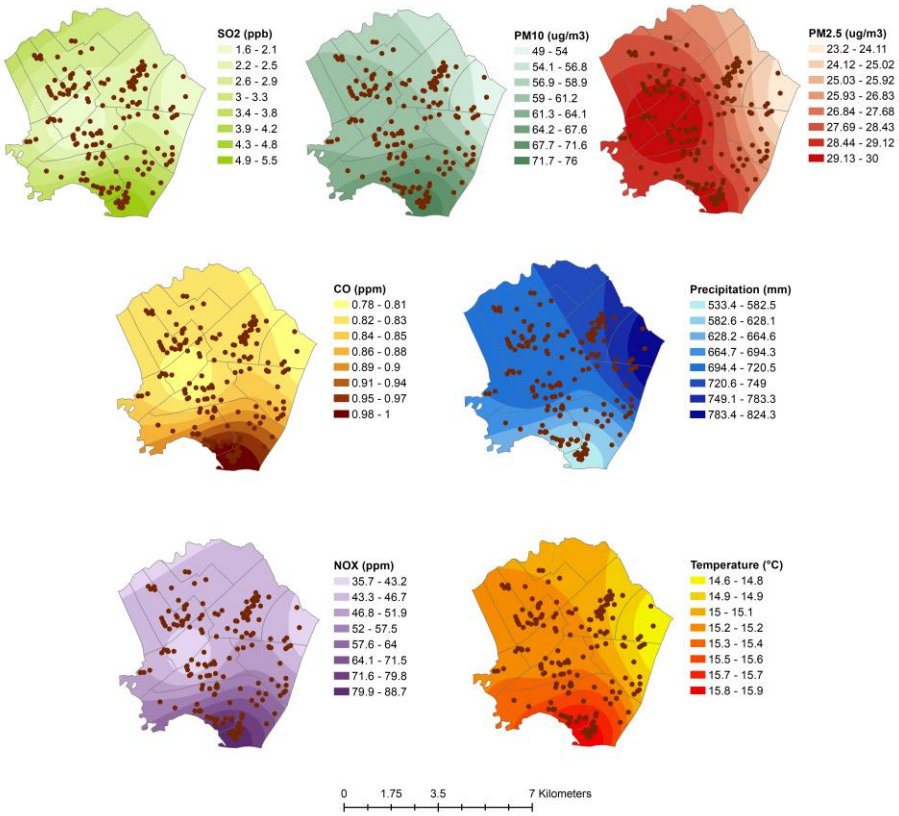
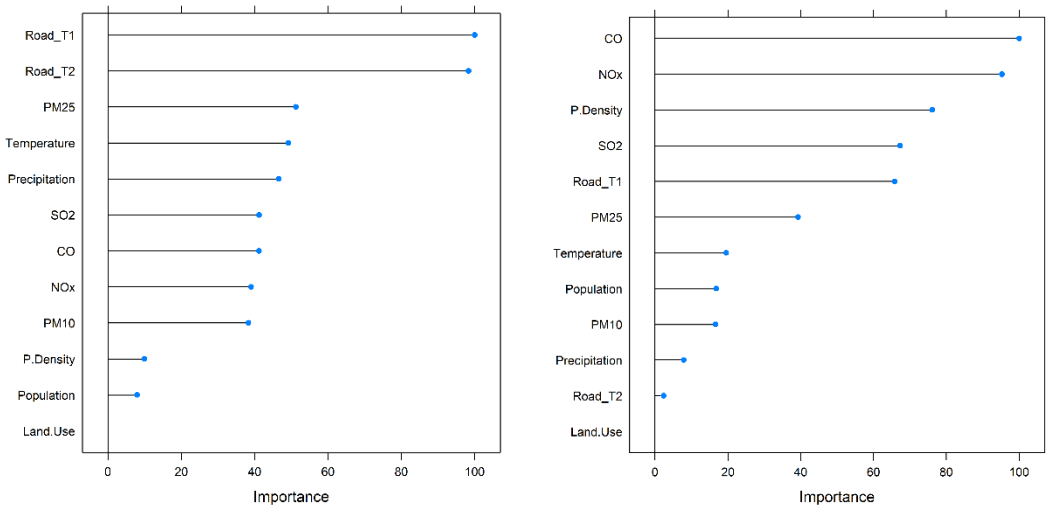


Fig. 6. Behavior of pollutants, meteorological variables, and field work results for 2016

Forecasting model with ML and GIS

Importance matrixes were created (see Fig. 7). The household proximity to roads (T1 and T2 in Fig. 7) variable had the strongest influence on the RF model, followed by temperature and $PM_{2,5}$. In the AdB model, household proximity to roads was the fifth most important variable. Variable behavior in the model is consistent with respect to the behavior recorded in 2016. The population-related variables are the least important in the RF model, while population density (P. Density in Fig. 7) plays an important role in the AdB model.



(a) (b)

Fig. 7. Hierarchization of predictor variables in the (a) RF and (b) AdB models

With respect to the models' performance, the RF model generated an AUC of 0.63 (see Fig. 8), in which the largest value was achieved through a model with 500 trees and 12 variables, which stabilized the error and prevented overfitting, resulting in an H measure of 0.10. The AdB model had an AUC of 0.52, for an H measure of 0.018.

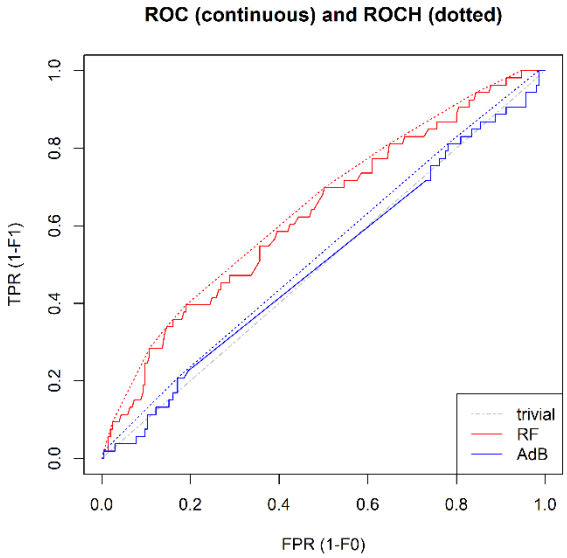


Fig. 8. ROC curve for RF and AdB

Forecasting zones with possible RD events

Based on the behavior of the variables introduced in the RF model, the most relevant zones in the locality related to exposed elements and external risk factors are Patio Bonito and Calandaima. Furthermore, considering the confluence of the most important variables in the RF model (proximity to T1 and T2 roads), the behavior of meteorological variables, and pollutants associated with both road quality and mobile source combustion ($PM_{2.5}$, CO, NO_x), there are hotspots present in each ZPU, which, depending on their intensity, enable the occurrence of possible RD cases (see Fig. 9).

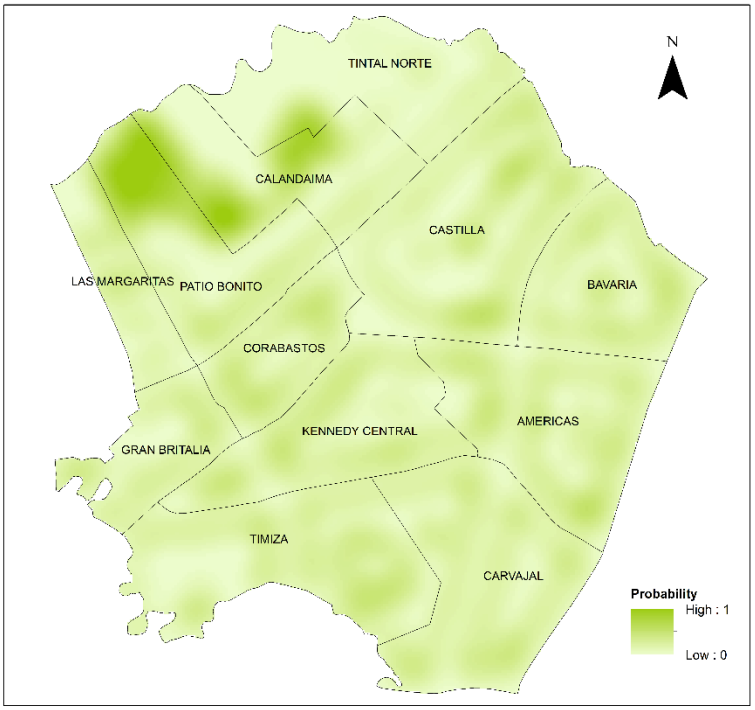


Fig. 9. Forecast of zones in which cases of RD could have occurred based on the RF model.

DISCUSSION

By applying ML tools, it was found that RF outperformed the AdB model for the H-measure, AUC, and accuracy (77.5% RF; 24.4% AdB). RF had a better odds ratio of 2.25, which reflects a sound diagnosis capacity and greater precision for the specific classification of possible cases of RD. In the same model, forecasting areas with high densities of possible cases of RD correlates with proximity to roads, $PM_{2.5}$, CO, SO_2 and meteorological variables. The high intensity classification of areas in Patio Bonito, Calandaima and specific points in Carvajal, Timiza, Kennedy Central and Castilla are supported primarily by the population's exposure to risk factors, including primary roadways on which different mobile sources of air pollution transit. However, there are other factors that influence RD events. For example, the low sensitivity of the model (11%) may be explained by the lack of variables that more accurately describe the behavior of the recorded events. This also responds to the influence exerted by important variables, as determined by their behavior.

It is worth noting that given their spatial behavior, in addition to identifying patterns of pollutant concentrations (Habibi and Alesheikh, 2017), or forecasting individual contaminants (Ivanov et al., 2018), it is necessary to consider the spatial behavior of combined risk factors and relate this behavior with the exposed population. This fact is sustained in the multiple pollutant exposure phenomena, which has not been addressed in many studies (Billionnet et al., 2012), as well as the proximity of the population to pollutants (Mazenq et al., 2017; Yu et al., 2019). This study gathered these experiences and made progress towards an innovative method of identifying zones where possible cases of RD may occur through the combined use of geostatistical tools and ML.

In this order of ideas, interpolation via IDW established the behavior of atmospheric pollutants and the meteorological variables, which were consolidated in the information bases for the correlation matrix and ML models. This aspect is consistent with prior experiences supported in studies developed by Gorai et al. (2018), Habibi and Alesheikh (2017), and Sajjadia et al. (2017), in which the transformation of information from observed points to continuous information was carried out to compare spatial behavior patterns. In different cases, atmospheric pollutant behavior is the primary variable for the analysis of its effects on a population's

health, which is consistent with this study that used ML and found that $PM_{2.5}$ was one of the most relevant variables.

Moreover, when information is collected in the field, conducting surveys has a related bias effect, such as selection bias. To reduce this effect, the survey was carried out in Kennedy's twelve ZPUs based on the sample, distributed by ZPU and land use. However, due to security concerns, entering some zones was difficult, which hindered the completion of the total number of surveys. This was the case in the Las Margaritas and Calandaima ZPUs. Furthermore, in the field data review process, it was found that due to spatial effects, some regions were not covered. As such, the number of surveys in these zones was increased to 912 for the final sample size value. Conducting a survey made it possible to identify possible respiratory system-related morbidity events in a spatial manner. Therefore, related biases may be limited in future uses with spatial location information that is recorded at health entities, and for security reasons, is not shared with the public.

Not having complete information of atmospheric pollutants and meteorological variables was one of the study's constraints. The air quality monitoring and tracking protocol (MAVDT, 2010) establishes a minimum data validity standard of 75%. That is, data whose information is at least 75% complete for the period analyzed is considered valid. Of the data used, an average of 78% met the temporal validity parameter. The NOx and SO_2 data represent 40% of the data that did not meet this requirement. $PM_{2.5}$ and CO had values of 38% and 14%, respectively. However, given the need to have information to feed the ML model and generate a weighted average of pollutant behavior and the meteorological variables for each year, the recorded information was evaluated in terms of its data trend in an analyzed time series based on the standard deviation of the variable for the pollutants that did not meet the required validity percentage in a given year and monitoring station. Thus, the data used exhibited a behavior recorded in its trend, which is largely in line with the required quality standard.

Detecting hotspots through spatial analysis with geostatistical and ML tools is useful to establish measures to reduce the vulnerability of people who are exposed to different health risk factors. Moreover, this approach facilitates the identification of important variables for the model, which is a prioritization tool. Nonetheless, due to different factors that influence people's health, the model could be strengthened through more available information to refine the characterization of the study area. This study's approach is useful as a support mechanism for urban planning projects, including the evaluation of territories' sustainable development performance. This approach could be applied in other fields to identify potential areas of interest, such as the agriculture sector to identify suitable soils, earth that is ready for the sowing of future crops, or to detect possible polluted soils due to different activities.

CONCLUSIONS

- This research developed a tool based on ML that presents the necessary stages to forecast hotspots in which possible RD cases may occur, based on the behavior of a territory's characterizing variables. Its application in a densely urban area is useful and replicable as it is a common characteristic in certain territories in developing countries. The micro-territorial nature of the study is relevant and innovative, as it differs from capital city and country approaches. This approach also enables researchers to generate useful technical support data for early warnings and contingency plans to mitigate impacts on air quality and population health, which also influences territories' economies.

- Using open-source software such as R and spatialization by means of open-source ML codes makes this study an easily replicable tool. These tools become stronger as more specific and spatialized information becomes available, and their advantages strengthen environmental health governance by public entities and the academic sector.

- The level of importance of pollutants such as $PM_{2.5}$, CO, SO_2 and meteorological variables influences the ML model's behavior. Relevant variables regarding the characteristics of the study area include: high vehicle flow of fossil fuel-powered automobiles (which explains the level of importance of the $PM_{2.5}$, CO and SO_2 variables); non-standard operating conditions; deterioration of local roads with the consequent generation of resuspended material; and residential areas with high population densities that are grouped together, where mixed land uses are integrated with commercial, industrial and service provision activities.

It is necessary to continue carrying out detailed studies on the exposed population that observe factors such as dose, duration, form of contact, age, sex, diet, personal characteristics, lifestyle and health condition, in order to determine the relative risk and establish these factors' behavior in the study area. In this study, 21.4% of the individuals surveyed reported having been diagnosed with respiratory diseases, of which 14.3% were individuals over 60 years of age, 51.8% are working-aged individuals, and 49.2% of those surveyed stated that they had lived in the study area for more than 10 years, demonstrating that exposure time is another variable of interest. These indicative figures include the broad spectrum of respiratory diseases, from the common cold to chronic and acute respiratory system diseases.

- Different areas reflect the confluence of risk factors and exposed elements. As such, the RF model established that an area of great interest could be in the Patio Bonito and Calandaima ZPU's. However, the residential characteristics of the Timiza, Kennedy Central and Carvajal ZPUs draw attention to the exposed population. RF perform better in terms of a model driver (AUC: 0.63; H measure: 0.1; accuracy: 77.5%), meaning that the results generated by the RF model are more accurate than those generated by an AdB model. Similarly, it can be concluded that it is possible to replicate this model in other areas or municipalities, and its accuracy can be improved by introducing specific data on the location with the highest exposure of patients attended to in consultations, emergency room visits, and hospitalizations related to RD, as well as information on the explicative variables for the analysis period. The combination of the tools applied in this study together with a pollutant dispersion model could increase the AUC, as well as the model's classification metrics.

- Lastly, it must be stressed that sustainable development refers to an increase in quality of life, through the interaction of social, environmental, and economic dimensions for equitable, livable, and viable development. A model of these characteristics becomes a preventive tool, which can contribute to reducing costs by addressing events associated with air pollution. As a territorial planning component, determining the influence of air pollution on a territory's sustainability can contribute to implementing policies instituted in the international framework. As air pollution increases, so does the number of workdays lost, reducing productivity. A better understanding of this phenomenon could contribute to zonal planning and determining the territorial organization of each zone.

Acknowledgments: Many thanks to the members of the Intelligence and Territorial Analysis Group of the Universidad Santo Tomás for their collaboration in conducting the fieldwork.

REFERENCES

Altman Douglas G, Bland J. Martin (1994) Diagnostic Tests 3: Receiver Operating Characteristic Plots. BMJ 309 (6948):188. https://doi.org/10.1136/bmj.309.6948.188

Billionnet C, Sherrill D, Annesi-Maesano I (2012) Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol 22:126–141. https://doi.org/10.1016/J.ANNEPIDEM.2011.11.004

Bobb JF, Valeri L, Claus Henn B, et al (2015) Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16:493–508. https://doi.org/10.1093/biostatistics/kxu058

Borja-Aburto VH (2000) Ecological studies. Salud Pública de México 42:533–538.

Breiman L (2001) Random Forests. Machine Learning 45:5–32. https://doi.org/10.1023/A:1010933404324

CRAN Comprensive R Archive Network (2018) R-3.5.2 for Windows (32/64 bit). https://cran.r-project.org/bin/windows/base/old/3.5.2/. Accessed 10 June 2019

DANE National Administrative Department of Statistics (2018) Multi-Purpose Survey -MS 2017. Bogotá, Colombia.

Del Valle Benavides AR (2017) ROC curves (Receiver-Operating-Characteristic) and their applications. Universidad de Sevilla.

DHS (2019) SALUDATA- Health Observatory of Bogota http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/salud-ambiental/consultaurgencias14anios/. Accessed 11 April 2019

Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM_{10} and $PM_{2.5}$ concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. Atmos Pollut Res 9:912-922. https://doi.org/10.1016/j.apr.2018.02.006

Galindo WG (2013) Construction dynamics by use, the locality of Kennedy 2002/2012. Bogotá.

García-Ubaque JC, García-Ubaque CA, Vaca-Bohórquez ML (2011) Medical consultation in productive age population related with air pollution levels in Bogota city. Procedia Environ Sci 4: 165–169. https://doi.org/10.1016/j.proenv.2011.03.020

Gorai AK, Tchounwou PB, Biswal S, et al (2018) Spatio-Temporal Variation of Particulate Matter ($PM_{2.5}$) Concentrations and its health impacts in a mega city, Delhi in India. Environ Health Insights 12:1-9. https://doi.org/10.1177/1178630218792861

Habibi R, Alesheikh AA, Mohammadinia A, et al (2017) An assessment of spatial pattern characterization of air pollution: A case study of CO and $PM_{2.5}$ in Tehran, Iran. ISPRS Int J Geo-Inf 6:270. https://doi.org/10.3390/ijgi6090270

Hand DJ (2009) Measuring classifier performance: A coherent alternative to the area under the ROC curve. Mach Learn 77:103–123. https://doi.org/10.1007/s10994-009-5119-5

Hernández B, Velasco-Mondragón HE (2000) Cross-sectional surveys. Salud Pública de México 42: 447–455

Huang K, Xiao Q, Meng X, et al (2018) Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China plain. Environ Pollut 242:675–683. https://doi.org/10.1016/j.envpol.2018.07.016

IDEAM Institute of Hydrology,Meteorology and Environmental Studies (2016) State of air quality in Colombia, 2011 – 2015 Report. Bogotá D.C.

Ivanov A, Voynikova D, Stoimenova M, et al (2018) Random forests models of particulate matter PM_{10}: A case study, in: AIP Conference Proceedings 2025, 030001. https://doi.org/10.1063/1.5064879

Jenks, George F (1967) The Data Model Concept in Statistical Mapping. International Yearbook of Cartography 7: 186-190

Kami JA (2019) A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions. Sci Total Environ 651:475–483. https://doi.org/10.1016/j.scitotenv.2018.09.196

Kassomenos P, Petrakis M, Sarigiannis D, et al (2011) Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model. Air Qual Atmos Health 4:263–272. https://doi.org/10.1007/s11869-011-0139-2

Kestenbaum B (2019) Epidemiology and Biostatistics. Seattle, USA. https://doi.org/10.1007/978-3-319-96644-1

Kuhn, Max, Kjell Johnson (2016) Applied Predictive Modeling. New York, USA. https://doi.org/10.1007/978-1-4614-6849-3

Lazcano-Ponce E, Fernández E, Salazar-Martínez E, et al (2000) Cohort studies. Methodology, biases and application. Salud Pública de México 42:230–241

Li S, Batterman S, Wasilevich E, et al (2011) Asthma exacerbation and proximity of residence to major roads: a population-based matched case-control study among the pediatric Medicaid population in Detroit, Michigan. Environ Health 10:34. https://doi.org/10.1186/1476-069X-10-34

Li Jin, Heap Andrew D (2014) Spatial interpolation methods applied in the environmental sciences: A review. Environ Modell Software 53:173-189. http://dx.doi.org/10.1016/j.envsoft.2013.12.008

Ly S, Charles C, Degr A (2011) Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. Hydrol Earth Syst Sci 15:2259-2274. https://doi.org/10.5194/hess-15-2259-2011

MAVDT Ministry of Environment, Housing and Territorial Development (2010) Protocol for air quality monitoring. Bogota, Colombia.

Mazenq J, Dubus J-C, Gaudart J, et al (2017) City housing atmospheric pollutant impact on emergency visit for asthma: A classification and regression tree approach. Respir Med 132:1–8. https://doi.org/10.1016/j.rmed.2017.09.004

Pandey G, Zhang B, Jian L (2013) Predicting submicron air pollution indicators: a machine learning approach. Environ Sci Processes Impacts. 15:996–1005. https://doi.org/10.1039/c3em30890a

Polezer G, Tadano YS, Siqueira HV, et al (2018) Assessing the impact of $PM_{2.5}$ on respiratory disease using artificial neural networks. Environ Pollut 235:394-403. https://doi.org/10.1016/j.envpol.2017.12.111

Ramírez O, Sánchez de la Campa AM, Amato F, et al (2018) Chemical composition and source apportionment of PM_{10} at an urban background site in a high–altitude Latin American megacity (Bogota, Colombia). Environ Pollut 233:142–155. https://doi.org/10.1016/j.envpol.2017.10.045

Reid CE, Jerrett M, Tager IB, et al (2016) Differential respiratory health effects from the 2008 northern California wildfires: A spatiotemporal approach. Environ Res 150:227–235. https://doi.org/10.1016/J.ENVRES.2016.06.012

Rodríguez-Villamizar LA, Rojas-Roa NY, Blanco-Becerra LC, et al (2018) Short-Term effects of air pollution on respiratory and circulatory morbidity in Colombia 2011−2014: A multi-city, time-series analysis. Int J Environ Res Public Health 15:2-12. https://doi.org/10.3390/ijerph15081610

Rokach, Lior, and Oded Maimon (2015) Data Mining with Decision Trees: Theory and Applications. 2nd ed. Singapore: World Scientific Publishing Co. Pte. Ltd. 5.

Salam MT, Islam T, Gilliland FD (2008) Recent evidence for adverse effects of residential proximity to traffic sources on asthma. Curr Opin Pulm Med 14:3–8. https://doi.org/10.1097/MCP.0b013e3282f1987a

Sajjadia SA, Zolfagharib G, Adabc H, et al (2017) Measurement and modeling of particulate matter concentrations: Applying spatial analysis and regression techniques to assess air quality. MethodsX 4:372–390. https://doi.org/10.1016/j.mex.2017.09.006

Schapire RE, Freund Y (2012) Boosting: foundations and algorithms, Adaptive computation and machine learning. MIT Press, London.

SDA District Secretariat for the Environment (2017) Air quality annual report of Bogota, 2016. Bogotá, Colombia.

SDP District Planning Secretariat (2018) Monograph 2017 Assessment of the main territorial, infrastructure, demographic and socio-economic aspects of the locality of Kennedy 08. Bogotá, Colombia.

Weizhen H, Zhengqiang L, Yuhuan Z, et al (2014) Using support vector regression to predict PM_{10} and $PM_{2.5}$, in: IOP Conference Series: Earth and Environmental Science. IOP. https://doi.org/10.1088/1755-1315/17/1/012268

Westerlund J, Urbain JP, Bonilla J (2014) Application of air quality combination forecasting to Bogota. Atmos Environ 89:22-28. https://doi.org/10.1016/j.atmosenv.2014.02.015

WHO (2006) WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. Geneva, Switzerland.

Yu Y, Yao S, Dong H, et al (2019) Association between short-term exposure to particulate matter air pollution and cause-specific mortality in Changzhou, China. Environ Res 170:7–15. https://doi.org/10.1016/j.envres.2018.11.041

Zhan Y, Luo Y, Deng X, et al (2017) Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm. Environ Pollut 233:464-473. https://doi.org/10.1016/j.atmosenv.2017.02.023

Apéndice B.

Using machine learning tools to classify sustainability levels in the development of urban ecosystems

Documento versión del autor publicado en la Revista: Sustainability (Electronic ISSN 2071-1050). Recibido el 31 de Marzo de 2020; revisado el 11 de Abril de 2020; aceptado el 16 de Abril de 2020; publicado el 20 de Abril de 2020. DOI: https://doi.org/10.3390/su12083326

Nidia Isabel Molina-Gómez[1,2], Karen Rodríguez-Rojas[1], Dayam Calderón-Rivera[1], José Luis Díaz[3] and P. Amparo López-Jiménez[2]

[1] Department of Environmental Engineering, Universidad Santo Tomás, Bogotá D.C., Colombia; nidiamolina@usantotomas.edu.co
[2] Department of Hydraulic and Environmental Engineering, Universitat Politècnica de València, Valencia., España; palopez@gmmf.upv.es
[3] Department of Civil and Agricultural Engineering, Universidad Nacional de Colombia, Bogotá D.C., Colombia; jodiazar@unal.edu.co

---

La identificación de un set de indicadores en el marco de la Agenda 2030 de los ODS fue el punto de partida para la evaluación del desempeño sostenible de la zona de estudio; además fue posible determinar las etiquetas de clasificación para pronosticar niveles de sostenibilidad. Este trabajo generó los insumos y el punto de partida para el diseño de la metodología para la determinación de la influencia de la calidad del aire en el desarrollo urbano sostenible. Se establecieron las variables de importancia en la clasificación de los niveles de avance en el desarrollo sostenible

---

ABSTRACT

Different studies have been carried out to evaluate the progress made by countries and cities towards achieving sustainability to compare its evolution. However, the micro-territorial level, which encompasses a community perspective, has not been examined through a comprehensive forecasting method of sustainability categories with machine learning tools. This study aims to establish a method to forecast sustainability levels of an urban ecosystem through supervised modeling. To this end, it was necessary to establish a set of indicators that characterize the dimensions of sustainable development, consistent with the Sustainable Development Goals. Using the data normalization technique to process the information and combining it in different dimensions made possible to identify the sustainability level of the urban zone for each year from 2009 to 2017. The resulting information was the basis for the supervised classification. It was found that the sustainability level in the micro-territory has been improving: from a low level in 2009, which increased to a medium level in the subsequent years. Forecasts of the sustainability levels of the zone were possible by using decision trees, neural networks and support vector machines, in which 70% of the data was used to train the machine learning tools, with the remaining 30% used for validation. According to the performance metrics, decision trees outperformed the other two tools.

Keywords: urban sustainability; indicators; supervised classification; micro-territories

1. Introduction

For decades, sustainable development has been a significant challenge for nations which is supported by, among other aspects, the environmental and socio-economic impacts associated with registered population growth. In 2018, 55% of the world's population lived in urban areas, which is expected to increase to 68% by 2050 [1]. The primary objective in addressing this challenge is to provide an orientation for a sustained improvement in the population's living conditions, which faces poverty, disease (associated with environmental and social determinants) and violence, among other situations. In this regard, the development and implementation of the Millennium Development Goals (MDGs) and the subsequent Sustainable Development Goals (SDGs) play an important role in determining the progress made towards achieving sustainable development.

The concept has been analyzed in different studies from different approaches [2–4], based on a broad spectrum of interpretations, primarily founded on the notion established in the report Our Common Future, which states "development that meets the needs of the present while considering the needs of future generations" [5] (p. 16). Notwithstanding the global nature of the term [2, 3], studies primarily focus on analyzing three fundamental pillars; environmental, social and economic dimensions. Each dimension has its own specific challenges with respect to territorial conditions, in addition to being connected and integrated with one another, in order to make sustainable development achievable.

Evaluating sustainability establishes a degree of development for urban ecosystems, in which natural and artificial structures interact and coexist. Ecosystem services, provided by natural systems, contribute to urban ecosystems' sustainability through the provision of goods and services. However, environmental conditions are altered (air emissions, waste, wastewater, among others) as the result of man-made structures and urban communities.

The pillars of sustainable development are looked at from a policy context, with a view towards an interaction between ecology and society; human ecology. The environmental dimension corresponds to natural resources and anthropogenic structures, while the biological community refers to the living components of ecosystems [6]. The social, economic and institutional dimensions are part of the social system that is modified by technological infrastructure, knowledge, and social organization.
Social systems' influence on ecosystem services impacts the environmental dimension, not just at the resources level, but also in its biological community. In this manner, these interactions have been measured through indicators, whose objective is to establish conditions for the analyzed resource, in order to make decisions about its resilience.

Population growth and the ensuing pressure on natural systems through the use and exploitation of resources, creates a need to understand these forms of pressure and possible measures that can be implemented to promote achieving the Sustainable Development Goals. Knowing the variation of sustainable development in a territory, based on its behavioral pattern, is an indispensable input for planning actions and measures. Natural ecosystems are basic to human life. As such, forecasting ecosystems' behavior, both natural and urban, can provide tools to protect human ecology.

Several studies have been developed to measure progress levels with respect to sustainability in countries and cities [2,4, 7–11], in addition to other studies which have created inputs for forecasting nations' sustainability levels by using machine learning tools [12–15]. These studies have established procedures for the calculation, aggregation and comparison of indicators in different settings, and have also proposed tools that can be useful in decision-making. However, these studies have been developed mainly from a global perspective, for a comparison between the behavior of countries and cities, leaving aside a more detailed territorial level approach, which is useful for the territorial synergy required to implement the SDGs. There is a need to integrate actors in an analyzed territory, in addition to structuring a comprehensive instrument to support the development of urban sustainability processes at the local level.

Machine learning tools have been used for decades in different settings to forecast future behavior of input information. With the generation of large1volumes of data, using these tools has become more useful in developing improvement strategies and analyzing sustainability from the smallest urban setting (organizations, households) to the territorial level. Therefore, it is essential to understand that achieving sustainable

development is not only carried out through a national policy perspective, but also in understanding the actions of territories that are part of cities and regions; urban micro-territories [16].

In this vein, this study seeks to establish a methodology for forecasting sustainability levels of an urban ecosystem through supervised modeling with machine learning tools. For the case study here described, the locality of Kennedy was selected, which is an urban territory in the city of Bogotá, the capital of Colombia. Kennedy has 1.2 million inhabitants with a rapidly growing population; 38% growth from 1993 to 2017. 5.3% of this population lives in multidimensional poverty, among which, the health dimension (60%) is where most people are affected [17]. Kennedy is characterized by being one of the most polluted zones in Bogotá in terms of air quality, in addition to having high levels of insecurity. Several economic and service activities with contrasting environmental, social and economic behavior interact in this urban micro-territory. The analysis period for this study was 2009 – 2017.

Developing the aspects contained herein is innovative in that it applies machine learning tools to a territorial analysis approach. This study analyzed the dimensions of sustainable development in a more specific territorial scope which addresses aspects such as the difficulty in accessing information, a common characteristic in Latin America. This study is pioneering as it not only includes opinions from experts and community residents in the territory, but also an analysis of complaints and requests in the context of urban needs. The territorial scope established for a sustainability analysis, in the field of human ecology, is a perspective that nations need to take into account in order to achieve better results related to sustainable development goals and targets. In general, there is a lack of machine learning models that forecast the sustainability behavior of urban territories, starting at the micro-territorial level, to support national and global perspectives for informed decision-making.

This study is structured as follows; following this introduction, a description is given of the different steps undertaken for the supervised modeling of sustainability levels. These include collecting information by evaluating sustainability levels, to applying machine learning tools, and an analysis of the same according to evaluation metrics. Afterwards, the results from applying this methodology in the case study are presented, in which the conditions of the micro-urban territory were identified, along with an indicator correlation within the framework of the sustainability dimensions. The territory's behavior over the years analyzed is presented through a categorization of sustainability levels, as well as the behavior of the machine learning models that were used. The study concludes with an analysis and discussion of the results, putting forth a suggested method to forecast sustainability levels in urban territories.

2. Materials and Methods

Several variables influence a territory's sustainability level, and their interaction affects its population's quality of life. Machine learning tools such as decision trees (DT), support vector machines (SVM) and artificial neural networks (ANN) were used in developing this study. A model to classify the sustainability levels of an urban area was created by applying these tools, which is useful for decision-making. This study consists of three relevant procedural paths: characterization of the study area with indicators, definition of the classification labels for the supervised learning model based on the calculation from the sustainable development index (SDI), and the development of machine learning models. The above made it possible to not only by creating a method, but also a model for the SDI classification of an urban area at the micro-territorial level. These stages were developed in a sequential manner as described below (see Figure 1).

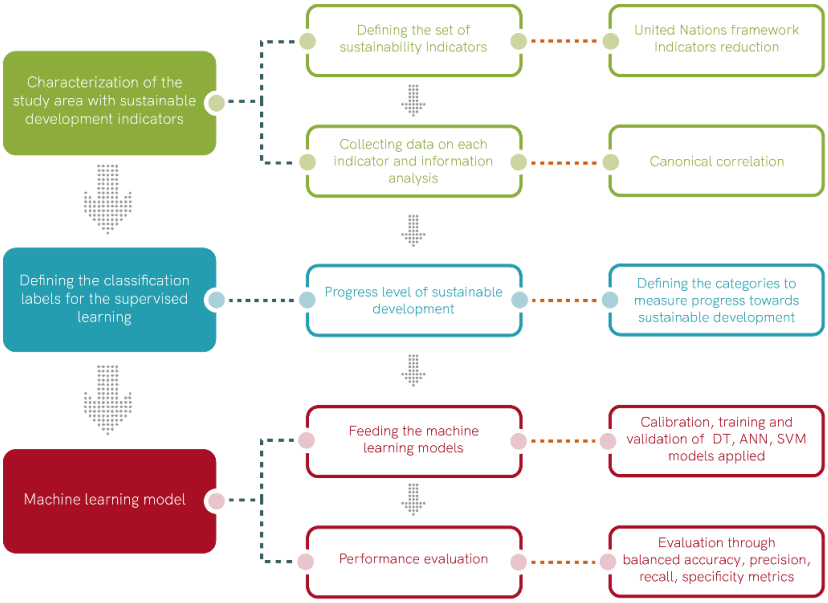


Figure 1 Methodological framework for the sustainability levels classified trough machine learning tools

2.1 Characterization of the study area with sustainable development indicators

The study area is the locality of Kennedy, an urban territory in Bogotá, the capital of Colombia. It is located at the coordinates 4°38′37″N 74°09′12″W (see Figure 2). This zone has a population density of 33,500 inhabitants/km^2, 36.7% greater than that of the city [17]. The locality is characterized by the presence of economic activities that include the provision of services, trade and certain manufacturing activities. 58.2% of the study area is residential, in addition to areas that are used for mixed purposes (services and trade). The zone has limited green space (6 m^2/inhabitant), with 3380 trees/km^2 [17]. Kennedy is made up of 12 zonal planning units in which different economic activities are developed along with housing areas.

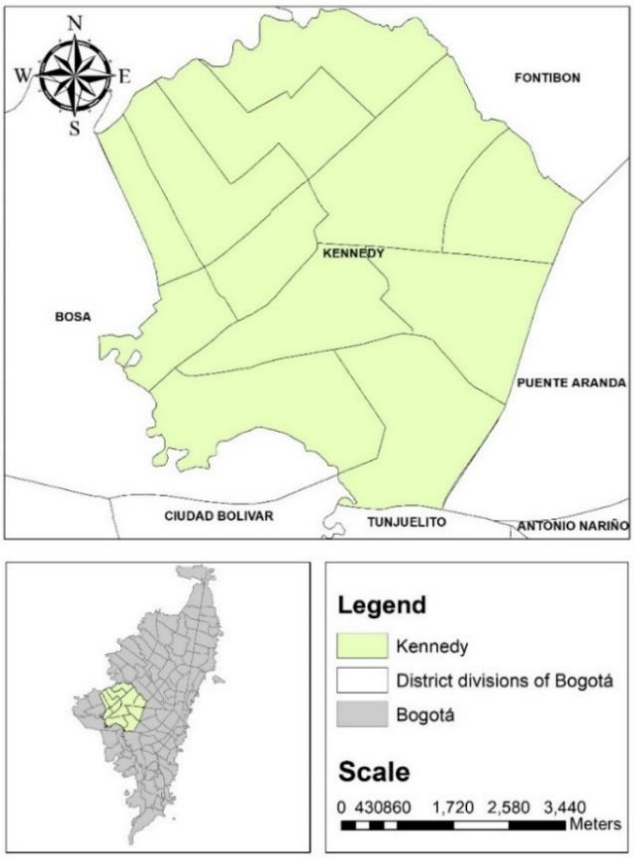


Figure 2 Locality of Kennedy as the study area in Bogota

2.1.1 Defining the set of sustainability indicators

To characterize this urban area, first, a set of environmental, social, economic and institutional indicators was established based on the framework put forth by the United Nations [18, 19]. This was followed by examining different studies that include analyses of sustainability dimension indicators [7, 9, 13, 20–26]. The steps taken above made it possible to identify a set of indicators capable of rating the progress level, in terms of sustainable development, of the urban area from 2009 to 2017. In selecting the indicators, consideration was also given to whether these were part of the goals, targets and indicators of the Sustainable Development Goals/Millennium Development Goals (SDG/MDG) [18, 27].

Subsequently, the indicators were reduced considering the following factors: a) the opinions from community residents regarding different subjects of interest, by analyzing complaints filed with public sector entities; b) the indicators' qualification characteristics; and c) the importance of the indicators according to criteria established by technical experts and people with an extensive knowledge of the territory.

The examination of the complaints filed by community members was carried out through a systematic frequency analysis. With respect to the indicators' qualification characteristics, eight characteristics were selected from the different studies analyzed [8, 20–22, 27–32]. These characteristics were: access to information, analytical soundness, universality, policy relevant and usefulness to users, use of a multidimensional approach, measurable, unambiguous, and systematic. Each one was rated on a 1 – 10 scale, in which 10 is the highest value of the indicator characteristic. Characteristics with a sum total less than 50 were discarded from the general base indicators.

In addition to analyzing the characteristics, a variation of the Delphi method was conducted [47], in which technical experts and people with an extensive knowledge of the territory evaluated the established importance

of the indicators. To this end, a web consultation was carried out using an electronic form addressed to technical experts in the specific areas of the indicators analyzed.
Furthermore, two workshops were held with the experts in the study area, to identify the importance of the indicators to residents in the territory.

The online expert consultation consisted of a series of closed-ended questions in which the participants stated their level of satisfaction with the eight characteristics for each indicator. A question was also included that inquired about the numerical importance of the indicator to achieve sustainable development in the study area.

With respect to the two workshops held with experts in the territory, a presentation was given on the project and how it was related to the SDGs, which was followed by an analysis of the indicators' importance in the territory. This evaluation was carried out through working groups and used a rating scale from 0 (low importance) to 5 (highly important).

It is important to note that considering the current participation spaces promoted by the local administration (i.e., local environment commission and economic observatory) 32 representatives from district entities, community leaders, and delegates from universities within the territory attended the workshops. This structured work developed so that different participants, with knowledgeable of the territory's priorities, could establish the importance of the indicators in the study area, making it possible to determine the set of indicators to evaluate the sustainability level of the urban zone.

2.1.2 Collecting data on each indicator and information analysis

The next step consisted of collecting information on each indicator for the studied period 2009 – 2017. Written reports from twenty-seven district entities were consulted, as well as information from technical documents and annual reports on the study area created by these entities and other institutions [17, 33–46].

A trend and behavioral analysis of the annual information for the period 2009–2017 was carried out for each indicator, which found incomplete information in some cases (8% of the total indicators used for the study period). Therefore, it was necessary to impute missing data in those cases in which specific annual information for the indicator was not available. The procedure followed in each case was to examine the indicator's behavior, and based on the same, the arithmetic average was taken by presenting an increasing or decreasing trend of the yearly information, or the moving average by presenting the variable's behavior from data with no apparent trend.

Once the set of annual indicators was established, a paired correlation analysis was performed via a canonical correlation analysis. A comparison was made of the linear behavior between the variables representing the environmental, social, economic and institutional dimensions. This procedure made possible to determine the canonical variables and their correlation level.

It is important to note that the sustainable development of a territory implies the integration of the environmental, social and economic pillars under the line of action defined through the institutional dimension. This study considered a characteristic parameter called "habitability," which is reflected in the relation between the environmental and social dimensions [7], in which indicators that describe the environmental health of a territory are relevant. An analysis was also carried out on the "viability" characteristic of the territory, which is based on the interaction between the environmental and economic dimensions [7], and also describes eco-efficiency indicators. Lastly, the analysis considered the importance of equitable development, based on the interaction between the social and economic dimensions [7], described by the indicators' relation within the framework of social efficiency. The institutional pillar was analyzed from a global perspective that provides the basis to develop the individual pillars and their interactions.

2.2. Progress level of sustainable development

In terms of planning, the term sustainable development has established a guideline from a global perspective, which aims to reduce inequities and improve conditions in the social, environmental and economic dimensions

with support from institutions. This study evaluated the study area's level of progress towards sustainability by considering the different indicators chosen for each pillar.

To calculate the sustainability level or sustainability development index (SDI), equation 1 (see Table 1) was used, in which the SDI is evaluated as the average behavior of the sub-indices for the environmental, social, economic and institutional pillars. Different indicators were established to be used as inputs for the process. These will be described in Section 3.

Each sub-index is calculated from the sum of the normalized indicators for each dimension, by considering the relative weight of each within the dimensional index (see equations 2 – 4 in Table 1).

Table 1. Equations used to calculate the sustainability level

Equation		Variables	Studies consulted
(1)	$SDI = (\frac{1}{4}) \sum_{1}^{4} DI$	*SDI = Sustainable development index* *DI = Indexes of each dimension*	[10, 47]
(2)	$I = \sum_{i=x}^{n} (w_i * x_i)$	*I = Indicator* *w_i = Relative weight of the indicator* *x_i = Normalized value of each indicator*	[10]
(3)	$y_t^i = \frac{x_t^i - \min(x_t^i)}{\max(x_t^i) - \min(x_t^i)} \in (0,1)$	*y_t^i = Normalized value* *x_t^i = recorded data value for period t* *min (x_t^i) = minimum data value of the indicator* *max (x_t^i) = maximum data value of the indicator*	[10, 48]
(4)	$DI = \sum_{1}^{n} (I)$	*DI = Index by dimension* *n = number of indicators of the dimension* *I = Indicator*	[47]

The relative weight, wi, in equation 2 was calculated by using the analytic hierarchy process (AHP). This process is based on a paired comparison of variables, considering the Saaty Rating Scale, 1987 [49], the eight defined characteristics, and the values of importance established in the participatory work developed with the technical experts and the people with extensive knowledge of the territory. It is noteworthy that the indicators' level of importance, which was stated by the people with extensive knowledge of the territory, was one of the characteristics included in the AHP assessment.

In parallel, the min-max scaling method was used to normalize the indicators (see equation 3 in Table 1). This method uses the distance between the maximum and minimum values of the analyzed indicators, considering the data of each indicator in the analysis period (2009 – 2017). Consequently, the indicator values were set to values in the 0 – 1 range, in which 0 represents the worst indicator performance and 1 reflects the best performance [10, 48, 50].

Lastly, by conjugating the variables in equation 1, the SDI was calculated for each analysis period. This same procedure was applied to the regular values or pre-established permissible levels for each indicator, either at the national level or based on international guidelines. This was done in order to compare the results for the study area with values established at the national and/or international levels that are deemed desirable for each indicator.

The calculated SDI values were the basis for the classification labels chosen in the supervised learning models that were applied in this study. Three categories for the sustainability levels were considered; low (0.0-0.33), medium (0.34-0.66) and high (0.67-1.0).

2.3. Machine learning model

This study used three different supervised machine learning tools to classify sustainable development levels: decision trees (C.5.0Tree); artificial neural networks (perceptron algorithm); and support vector machines (SVMradial) were used.

Decision trees (DTs) are a hierarchical predictive model of decisions and their consequences. They consist of nodes, branches and leaves that characterize the model, and also establish the complexity of the decision tree. Complexity characteristics include the depth of the decision tree and the number of attributes used; the more complex the decision tree is, the more complexity there will be with respect to the accuracy of the results. Induction rules are applied when developing decision trees [51]. Different algorithms for decision trees have been developed, including the C.5.0tree, which evolved from C4.5. The C5.0tree algorithm is characterized by using entropy to measure the purity of tree divisions. This algorithm includes or removes predictors (in this case indicators) based on their relationship with the labels established for supervised learning. In this manner, the model that is created includes only the most important predictors, taking into consideration that the error rate is reduced. In the event that the error rate is higher, due to not having included all the predictors in the classification model, they are left as predictors for the model [52].

For their part, artificial neural networks (ANNs) are mathematical models inspired by the biological functioning of neurons [52]. As with decision trees, this model is composed of nodes. In this case, they act as input, output or intermediate processors connected to each other through links. They are characterized by their use of adaptive learning and self-organizing algorithms, and they process information in a non-linear manner. The node receives an input that has an associated weight, which is modified in the learning process. Basis and activation functions are necessary for the network to function.

Lastly, as a classification tool, support-vector machines (SVMs) use proximity to classify samples in a vector space. The maximum distance in the hyperplane is measured by the points closest to it. In this manner, the categories will have a distance from each side of the hyperplane, serving as a classification space. The representation by the mean of Kernel functions provides a solution to this problem; projecting the information to a larger characteristic space, which increases the computational capacity of the linear learning machine [53].

2.3.1 Information required to feed the models

The information used to feed the models corresponds to two important inputs; indicators according to dimension, and supervised classification parameters.

a) Indicators that describe the behavior of the study area according to the sustainable development dimension; environmental, social, economic and institutional. This study used machine learning tools on the indicators that were normalized through equation 3 in Table 1, with information on yearly (81 indicators) and monthly (16 indicators) scales. An annualized basis of indicators was used taking into account reporting characteristics in the study area. However, given the nature of how DTs, SVMs, and ANNs function, the results were derived from monthly information.

This study aims to establish a forecasting method for sustainability levels by using machine learning tools. Therefore, examples of variable data are required for the process to train and validate the models. Consequently, in the cases in which it was not possible to complete the monthly information, the indicator was discarded from the information base that would feed the model. Furthermore, in cases in which invariable information behavior was observed, these indicators were not included in the learning model with monthly information. That is, indicators such as the drinking water supply, which during the year does not vary significantly, but which over the years has a degree of variation, as well as wastewater treatment for example, were eliminated from the information set to be included in the model. In each case it was verified that the sustainability pillars were represented in the indicators, in order to develop the learning and classification process.

b) Regarding the selection of classification parameters for supervised modeling, the results from the evaluation on the study area's sustainability level were used to establish the supervised classification labels. Three sustainability level categories were established: high (0.67-1.0), medium (0.34-0.66) and low (0.0-0.33). It is

worth noting that given the characteristics of the results from the index calculation, scenarios were created to allow training data to be entered into the model, specifically for the low and high sustainability labels. These scenarios were generated by considering each indicator's threshold value, ensuring that the models had enough training examples in the data set and for validation, in accordance with the proposed scenarios. A 108-data point set was available for monthly reporting purposes, 70% of which was used for training and 30% for validation in the classification process. The same ratio for training and validating was applied to the yearly data set.

2.3.2 Performance evaluation of machine learning models

The metrics used in each model to measure its performance correspond to balanced accuracy, precision, recall, and specificity, or true negative rate, as determined by the confusion matrix. The matrix is a 3x3 table with different combinations of predicted and actual values regarding the classification labels (in this case a high, medium and low sustainability level). The balance accuracy metric prevents inflated performance estimates in unbalanced data sets. The metric determined the accuracy of the classifier to forecast each sustainability category: high, medium and low. In this vein, if the complete set of labels predicted for a sample strictly coincide with the real set of labels, the accuracy of the subset is 1.0. For its part, the precision metric made it possible to know the capacity of the classifier to not classify a result in a sustainability category or level that belongs to another category. The best results from this metric are 1.0, falling in an average close to 0.0. The recall metric refers to the classifier's capability to find all samples belonging to the sustainability category being evaluated, with a value of 1.0 referring to the best results for the metric.

Furthermore, the level of importance of the input variables was established by using the Gini index in the implementation of the supervised learning models.

To develop machine learning models, the open source R software was used along with the caret package library, specifically for the following models: decision tree (method: C5.0Tree) [54]; artificial neural networks (method and package: nnet) [55], and the function of the package e1071 for the support vector machine [56].

3. Results

3.1. Characterization of the study area

A set of 81 indicators was established to be used as inputs for the process. The table presented in the supplementary material (Appendix A) puts forth a description of the indicator set according to the dimension to which it belongs, the intersection if the indicator is part of an intersection (livable, equitable, viable), as well as the related sustainable development goal and target. Each indicator has an identification code, a combination of a letter and a number. The E letter identifies indicators belonging to the environmental dimension; the S letter identifies indicators belonging to the social dimension; the letters EC identify indicators belonging to the economic dimension and, the letter I identifies indicators of the institutional dimension. Table 2 presents an outline of the indicator set, displaying the number of indicators according to the characteristics established for each cell.

Table 2. Set of indicators for the analysis of urban sustainability in the micro-territory

Dimension	**Environmental = 13**	**Social = 47**	**Economic = 16**	**Institutional = 4**
Intersection	Livable = 20; Equitable = 24; Viable = 1; Sustainable = 14			
Subject	Air = 4 Water = 4 Waste = 2 Green spaces = 3	Health = 21 Education = 6 Demography = 3 Security = 4 Coverage of public services = 6 Transportation = 2	Economic structure = 7 Poverty = 3 Consumption and production = 7 Income and expenditure = 4 Employment = 4	Government = 2 Social community services = 2

Dimension	**Environmental = 13**	**Social = 47**	**Economic = 16**	**Institutional = 4**
Intersection	Livable = 20; Equitable = 24; Viable = 1; Sustainable = 14			
Related SDGs	▪ Clean water and sanitation ▪ Sustainable cities and communities ▪ Life of terrestrial ecosystems	▪ End of poverty ▪ Zero hunger ▪ Health & well-being ▪ Quality education ▪ Gender equality ▪ Clean water and sanitation ▪ Industry, innovation and infrastructure ▪ Sustainable cities and communities ▪ Peace, justice and solid institutions	▪ End of poverty ▪ Health & well-being ▪ Affordable and no polluting energy ▪ Decent work and economic growth ▪ Sustainable cities and communities	▪ End of poverty ▪ Partnerships to achieve goals

With regard to the environmental dimension, over the analysis period, the study zone has improved in terms of its indicators on air quality, waste collection and areas allocated for green spaces. However, domestic wastewater generated in the locality is discharged into water sources without any type of treatment. On the other hand, while some indicators behave in a relatively constant manner, the importance of their improvement is noteworthy, specifically km^2 of green areas and recreational spaces.

With respect to the social dimension, a substantial number of indicators (25%) are related to the subject of health, given the influence exercised by socio-environmental determinants. These indicators' behavior does not reflect a marked upward or downward trend but responds specifically to the health determinant conditions present each year in the study area. Despite the variability, improvements are seen in indicators such as the child malnutrition rate, under-five mortality rate, all-cause infant mortality rate, and maternal mortality ratio.

Regarding the education indicators, gross education coverage decreased in 2016 and 2017 in the study area. However, the indicator behavior improved for areas such as years of schooling completed, illiteracy rate, population with middle and high school level education, and school attendance rate during the analysis period. Furthermore, with respect to population, the number of inhabitants per square kilometer has seen an upward trend, but the number of square kilometers with informal settlements has decreased, while coverage of the storm drainage system and the number of passengers transported by the mass transportation system have increased.

The study area is noted for having many security concerns, shown in indicators such as theft, aggravated robbery and reports of domestic, family and child abuse; indicators which had a negative behavior trend during the study period.

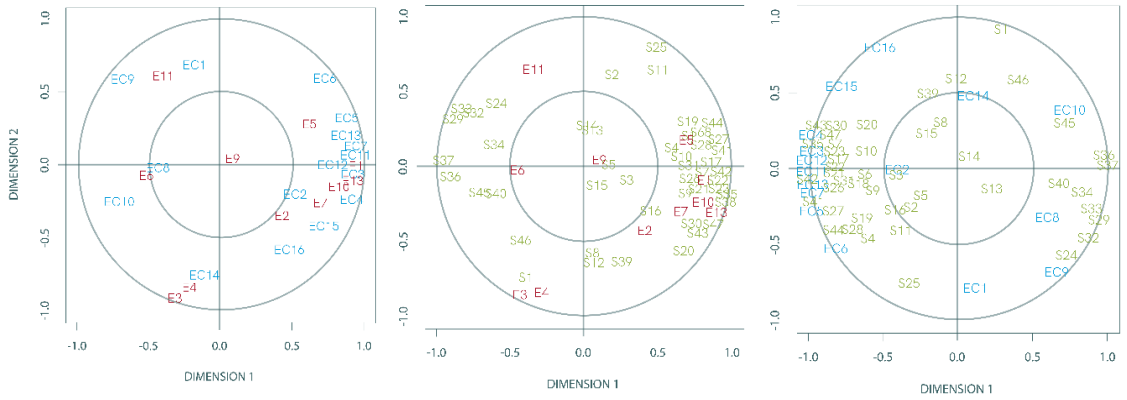
Concerning its economic structure, the locality has high levels of its population living under the poverty line, with its highest recorded value in 2015, with 183,966 inhabitants in this condition. In the final two years of the study period, this indicator decreased by nearly 10%, in which there was a higher risk of water shortages (on average 171 people ± 42). However, there was an improvement in indicators such as access to electricity (yearly increase of nearly 2%), per capita household income, and improvements to the road network in the urban area.

Lastly, the institutional dimension is supported by policies and actions from the institutional sphere to meet the needs of the other pillars. The indicators that comprise this dimension had stable behavior during the analysis period.

As shown by the indicators, these characteristics are consistent with the frequency analysis of complaints filed by community members, which had high values concerning safety (15% of the 46,800 written complaints analyzed). This is in addition to the situation of the canonical correlation that enabled the indicators to be conjugated, which is described below.

3.1.1 Canonical correlation

In the correlation analysis of the 81 indicators with an annual frequency in the period 2009 – 2017, the comparison between environmental protection and economic growth (see Figure 3) found a relation between indicators such as PM10, PM2.5, access to public services and the unemployment rate. The upper right-hand margin of Figure 3 shows an important grouping of economic indicators. All have a positive behavior, in the sense of increased per capita household income (EC5), an increase in energy consumption (EC12), and growth of the employed population (EC3), for example. In this grouping, there are environmental indicators such as the average annual concentration of PM10 (E1), number of trees per hectare (E13), and the water quality of the Tunjuelito River (E10). Furthermore, the same quadrant includes indicators regarding PM2.5 (E2) and the road network in good condition (EC15), both with improving trends.



(a) **(b)** **(c)**

Figure 3. Canonical correlation between indicators of the a) Environmental and economic pillars; b) Environmental and social pillars; c) Social- economic pillars. Indicators with an E code refer to environmental indicators; indicators with an EC code refer to economic indicators; and indicators with a S code are social indicators. In Figures 3a, 3b, and 3c, Dimensions 1 and 2 are the canonical variables that make the best representation of the total variation of indicators interactions. These canonical variables maximize the discrimination between groups of indicators. Appendix A shows the environmental, social, and economic indicators, represented in this figure through letters E, EC, or S, followed by an identification number.

The second chart (Figure 3b) shows an initial grouping of indicators that measure mortality rates: all-cause infant mortality (S6), under-five mortality from pneumonia (S4), under-five mortality (S10), perinatal mortality (S18), and life expectancy at birth (S28). The air quality index (E5) is included within this set of indicators in Figure 3b. There is also a set of health indicators such as acute malnutrition in children under 5 (S7) and the infant death rate (S21); indicators that characterize the physical conditions of the study area such as km2 of areas susceptible to flooding (S38), as well as service indicators, which include the number of passengers who commute via the mass transportation system (S35) and households with access to natural gas service (S42). Furthermore, there are education indicators such as school attendance rate (S23), average years of schooling completed (S22), and population with a middle and high school education (S26). Another social indicator in this grouping corresponds to deaths due to firearms (S31). In addition to this set, there is the average annual concentration of PM10 (E1) and closely related indicators such as the water quality of the Tunjuelito River (E10) and the number of trees per hectare (E13). This same chart shows the closeness of indicators that report excesses of PM10 (E3) and PM2.5 (E4), as well as the indicator that corresponds to the mortality rate due to cardiopulmonary disease, pulmonary circulation diseases and other forms of heart disease (S1).

Lastly, the third graph (see Figure 3c) shows a comparison between social inclusion and economic growth, in which there is a correlation between indicators such as access to public services, the economically active population, and education level.

3.2. Progress level of sustainable development

Applying equations 1 to 4 (see Table 1) the sustainability categories were calculated for each analysis year in Kennedy. The locality has had low to medium sustainability levels (see Figure 4). However, the behavior in 2016 and 2017 surpassed the medium sustainability level (0.33-0.66). Moreover, the biogram presented in Figure 5 shows the behavior of the environmental, social, economic and institutional sub-indices for the study area.

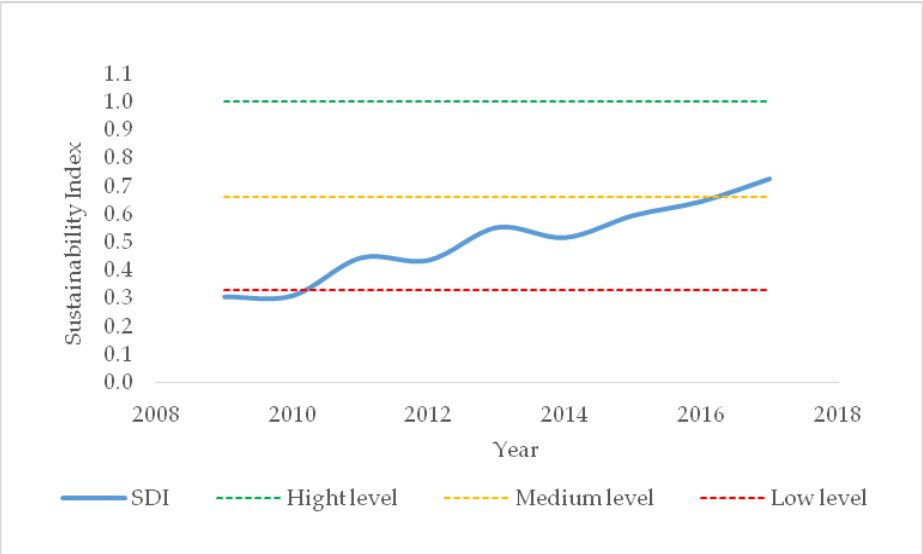


Figure 4. 2009 – 2017 sustainability index for the locality of Kennedy

Figure 5 shows the influence from the institutional and economic dimensions, with a lag seen in the environmental pillar when compared with the other dimensions. In general, the behavior related to the SDI has improved for each dimension from 2015 to 2017.

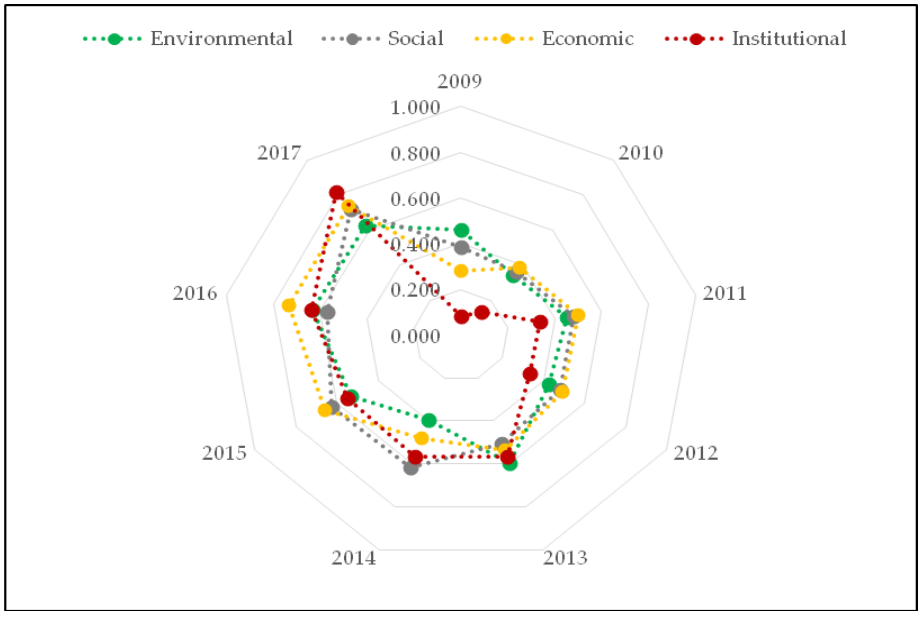


Figure 5. Biogram of the influence of the sustainable development dimensions for each year of the study period in the Kennedy urban area

3.3. Machine learning model

As mentioned in the methodological description, yearly and monthly information was used to develop the models. Each model was calibrated based on specific parameters for each machine learning tool, following the selection criteria provided by the kappa and accuracy measurements, as presented in Table 3.

Table 3. Calibration parameters for the machine learning tools to define the classification model of the Sustainable Development Index for the urban micro-territory.

Tool	Calibration parameters
Decision trees (C.5.0 Tree)	Iterations: 1 Kappa = 0.28±0.17; Accuracy: 0.72±0.11
Artificial neural networks	Size: 1 Weight decay: 0.1 Kappa = 0.42±0.12; Accuracy: 0.81±0.07
Support vector machine (SVMradial)	Sigma: 0.04 c: 1 Kappa = 0.4±0.22; Accuracy: 0.83±0.06

Applying the models found that due to the limited number of observations (9 data points for each indicator), models based on yearly information turn out to be inconclusive. Given the low volume of observations entered, it was not possible to forecast sustainability levels. However, using a monthly scale increased the number of observations, which enabled a greater volume of information to be available to train and validate the models. Table 4 presents the results for the three models developed. The labels high, medium and low correspond to the classification categories of the sustainability level assigned to the model for training and subsequent forecasting. Values with results in the 0.67 – 1 range belong to the high sustainability category; values with results in the 0.34 – 0.66 range correspond to the medium category, and values with results ranging from 0 to 0.33, belong to the low category.

Table 4. Metrics generated by the machine learning models in the classification of sustainability levels in the micro-territory

Model	**Balanced Accuracy**			**Precision**			**Recall**			**Specificity**		
	High	**Medium**	**Low**	**High**	**Medium**	**Low**	**High**	**Medium**	**Low**	**High**	**Medium**	**Low**
Decision tree - C.5.0 Tree	0.95	0.81	0.60	0.75	0.90	0.33	1.00	0.82	0.33	0.91	0.80	0.86
Neural networks - Nnet	0.96	0.80	0.50	0.75	0.85	-	1.00	1.00	0.00	0.92	0.60	1.00
Support Vector Machine - SVMradial	0.79	0.70	0.50	0.67	0.79	-	0.67	1.00	0.00	0.92	0.40	1.00

As this is a multi-class model, as a whole, the decision tree model yields the best metrics (see Table 4). Decision trees and neural networks were 95% and 96% accurate, respectively. The high and medium territory sustainability categories were 81% and 80% accurate, respectively. While the support vector machine was not as accurate, it performed well in the classification, with values of 79% for the high category and 70% for the medium category.

The accuracy of the low classification category indicate that neural networks and the support vector machine classify the information for this category in a random manner; only decision trees were 60% accurate in the low classification category. These values are consistent with the results established by the precision metric, in which the decision tree and neural network models correctly predicted 75% of the labels in the high category. According to the recall metric, 100% of the labels for this category were forecasted. With respect to the medium

sustainability category, the precision metric shows that 90% of the forecasted labels were correct in the decision tree model and according to the recall metric, 82% of the category was forecasted.

3.3.1. Variable importance based on the Gini index

For the decision tree model, the variables with the greatest importance were: population with access to health services (S47), residential per capita water consumption (EC16), and excess PM10 (E3) (see Figure 6). For the neural network model, the variables with the greatest importance were: reports of violence and domestic abuse (S32), excess PM10 (E3), theft and aggravated robbery (S33), mortality rate due to pneumonia in adults older than 64 years of age (S3), and average annual concentration of PM2.5 (E2) (see Figure 6). With respect to the support vector model, the most influential variables that exceeded 60% importance were: population with access to health services (S47), passengers who commute via the public mass transportation system (S35), reports of violence and domestic abuse (S32), energy consumption (EC13), average annual concentration of PM2.5 (E2), excess PM10 (E3), and residential per capita water consumption (EC16). The above can be seen in Figures 6a, 6b and 6c, related to each forecasted level of sustainable development.

When comparing the most influential variables in the models, the excess of PM10 variable (E3) is present in the three applied models, with similar levels of importance; 64% for ANN, 78.4% for SVM, and 37.8% for DT, for the high and medium sustainability categories (see Figures 6a & 6b). Additionally, its importance drops by 19 percentage points in the low category for the SVM model (see Figure 6c). While the population with access to health services variable (S47) is the most important variable in the DT and SVM models, it scores less than 30% in the ANN model. The role of the social dimension's variables, related to security, stands out, given its influence on the classification of sustainability levels of the urban area.

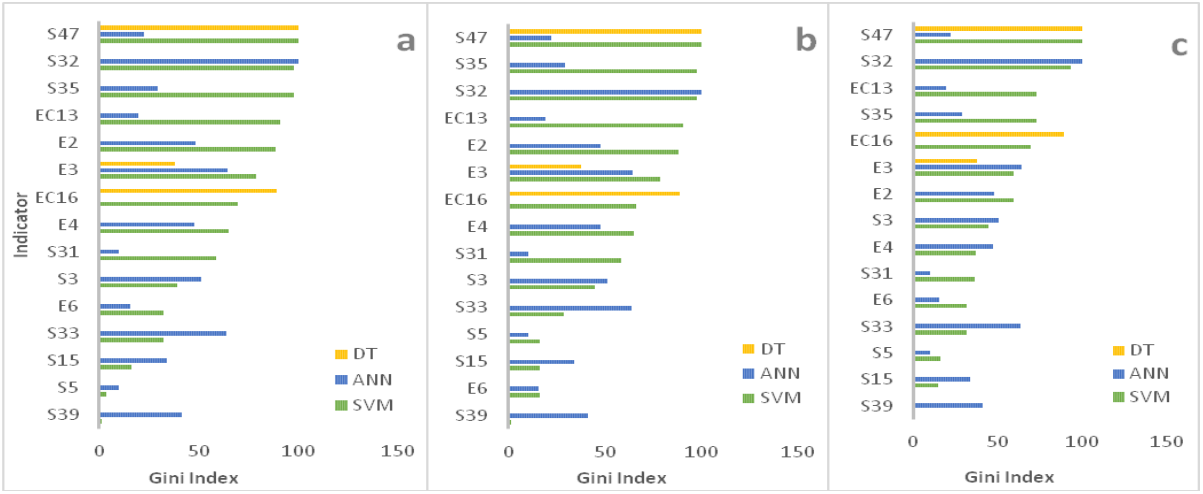


Figure 6. Variable importance according to the classification of the sustainable development index; **a**: high sustainable development index score; **b**: medium sustainable development index score; **c**: low sustainable development index score. DT: decision tree; ANN: artificial neural network; SVM: support vector machine. Appendix A includes the description of the environmental (E code), social (S code), and economic (EC code) indicators.

4. Discussion

The canonical correlation analysis found that the behavior described by the indicators shows that the urban area has different needs regarding the sustainability pillars and residents' quality of life. This is reflected in the interactions between indicators that seemingly do not show a direct relationship, yet describe specific determinants of the micro-territory's reality in the habitable and equitable interactions in the urban area [10].

There is an interaction between indicators such as the employed population between 12 and 64 years old (EC3), the economically active population (EC2), and indicators related to the habitable interaction, such as: water quality of the Tunjuelo River (E13) and trees per hectare (E10). In addition to the analysis, there is a connection between indicators regarding economic issues and those that address social characteristics in the area, in terms

of education and security (theft and violence). The grouping with the canonical correlation reflects behavior as described by Tanguay (2017) [10], for each of the pillars' interactions. Furthermore, the grouping of sustainability indicators such as passengers transported (S35), aging rate (S30), households with access to water (S42), energy consumption (EC12) and acute malnutrition of children (S7), which, despite the classification of specific issues, result in the interaction of sustainability dimensions in the territory. With respect to these interactions, it is important to note that the priorities in evaluating and measuring urban sustainability are determined by the territorial characteristics themselves [2]. That said, it is necessary to establish a comparison line in order to identify territories' evolution. To this end, the Sustainable Development Goals and its targets are an appropriate platform that brings together common goals.

Previous studies on the city of Bogotá have determined that the most relevant variables in the sustainable development index are poverty, crime and unemployment [4], in which the index was calculated by applying a sustainability assessment by fuzzy evaluation. These variables are consistent with the results from this study in the complaints analysis as an input to prioritize indicators and calculate the Sustainable Development Index. However, it is considered that they should not be the only factor of interest as sustainable development is achievable only to the extent that interactions are addressed and balanced, such as the livable, viable and equitable dimensions [7, 11], as shown by the canonical correlation analysis.

These indicators' behavior establishes that the population increase in the urban area and its resulting impacts, substantiate the need to advance a process of continuous feedback in order to support improving the conditions of the environmental, social, economic and institutional dimensions in territories. These are the results obtained from evaluating the Sustainable Development Index.

Kennedy is the second most populated territory in Bogotá. According to the SDI evaluation, the SDI of the urban area has moved from the low to the medium category over the period 2009 – 2015, with values that surpassed the medium sustainability category in 2016 and 2017 (See Figure 4). Prior studies have determined that Bogotá has reached a medium sustainability level (0.55, on a 0 – 1 scale), ranking 88 among 106 European, African, Asian and Latin America cities [4]. Another study that applied multivariate statistical techniques [8], identified a medium sustainability level for Kennedy. Despite the difference in the methods applied to evaluate sustainability, these studies were consistent with the results presented in this paper. Furthermore, the variation in the numerical values recorded is limited, which is counterbalanced by studies that analyzed the variation in results with respect to the methodological variation in calculating sustainability, which yielded similar results even with different methodologies applied [10]. That said, it is important to note the importance of indicator selection for a relevant evaluation of sustainability.

Furthermore, a comparison of the influence of a micro-territory with better socio-economic behavior than Kennedy found that the results obtained through the SDI evaluation for Kennedy in this study are consistent with results from prior studies [8]. Teusaquillo is another micro-territory in Bogota, which unlike Kennedy is characterized by having greater purchasing power, more employed people, as well as having better educational, financial, cultural, and recreational services. In this vein, according to Carrillo & Toca (2013) [8], Teusaquillo achieved a high sustainable level in the evaluation. These are aspects which, despite the difference in methodologies, influence territories' progress towards sustainability.

Moreover, it has been noted that the development and implementation of a machine learning model require enough observations to ensure adequate training and validation of its behavior. This project faced limitations associated with not having enough information; some of the available information corresponds to specific data concerning the city of Bogotá, primarily corresponding to the periods in which surveys, reports on the implementation of government plans, or the gathering of information for specific purposes were carried out. Planning and territorial evaluation processes do not consider creating range indicators for urban sustainability dimensions at the micro-urban territory level. In the face of these limitations, the following three specific aspects stand out:

1) Benchmarking was used to select the indicators for this study, which was carried out by examining many existing researches on these types of indicators, in addition to reviewing the framework of the SDGs to achieve congruity amongst the indicators. The analyses presented herein are consistent with those presented by L.-Y. Shen et al. (2011); Shen et al. (2013) and Verma et al. (2018), regarding the need to have valid objectives and

targets for each territory as a clear support mechanism to evaluate progress made towards sustainability [2, 3, 57]. The indicators are matters of governance, but not issued by the government [8]. As such, it is necessary to develop a collection of historical data on territorial behavior, as this provides evidence of territories' evolution and support for sustainable development processes. Furthermore, given that population is an essential component of urban activities [2], participation from interest groups and including their needs to determine the set of indicators is necessary.

2) The evolution of territories, as a goal of sustainable development, in which human beings are the central axis of governments, requires coherence and coordination to identify, collect and process information. Several studies use national statistics that have been published on various platforms for years prior to the implementation of the Millennium Development Goals as the basis for their information sources. Unfortunately, a clear example of the need to prioritize indicators can be seen in Latin American territories, where a greater impulse is required in information management, as demonstrated in the micro-territory analyzed in this study. It is also a mitigating circumstance for the capital city's position in the ranking of cities with the lowest sustainability levels, according to the results from Phillis et al. (2017) [4].

3) At the international level, proposals for forecasting sustainable development in different cities and countries have been developed using indicators with a yearly scale [12, 13, 15]. However, the present study was not able to yield conclusive results for this time scale. In applying DTs, as one of the simplest tools for this type of classification problem, and the SVM and ANNs as robust tools, nine observations were not enough to properly train the model and validate its results. As stated above, 70% of the data was used for training and 30% for behavioral validation. Therefore, using these types of tools requires large amounts of information, which prevents generalization problems and ensures the information's quality to support decision-making. In this vein, the model for this study reduced the working scale to monthly indicators, finding that the decision trees had the best behavior, with neural networks having potential for improvement.

Lastly, the method applied and structured through this study established a logical procedure that begins with identifying the most influential parameters in an urban territory and concludes with forecasting their behavior in terms of sustainable development (see Figure 7). This procedure collected experiences developed in various studies that combine community participation in the territory, the technical expertise of professionals in areas of sustainable development, and the robustness offered by machine learning tools such as decision trees, neural networks and support vector machines. This study was innovative in that it took a methodological step forward by integrating the community who are affected by their government's decisions, while including experiences from different studies, and the vision of the SDGs. It also integrated different tools for decision making, to be used for annual and statistical collection plans, as well as to manage the different resources that characterize the sustainability pillars.

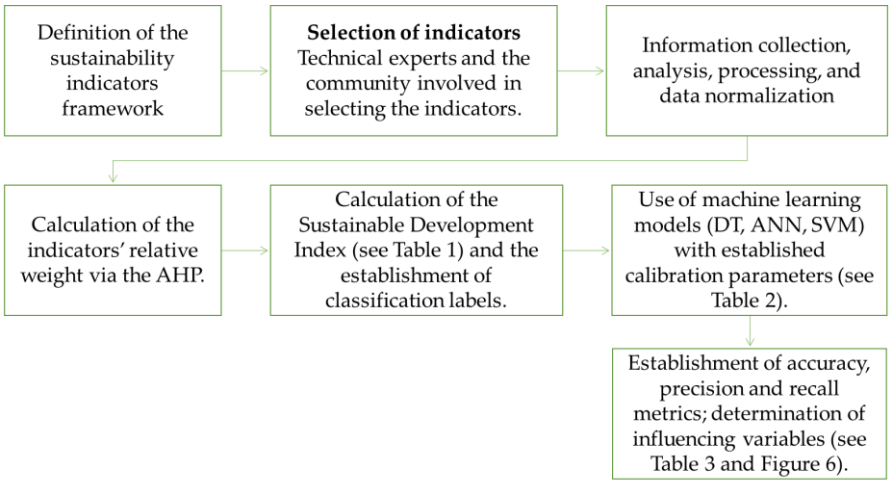


Figure 7. Suggested methodological process to classify the sustainable development index in urban micro-territories

Future studies should focus on the importance of having spatialized information, which enables the identification of the behavior of habitability interactions and the viability of sustainable development in different territories. This information can be used to forecast sustainability categories with machine learning tools as additional support for decision-making. Similarly, it can resolve difficulties in accessing information [2], even at the level of an urban micro-territory analysis, which was chosen for this study.

5. Conclusions

As shown in the present research, urban ecosystems include a combination of diverse micro-ecosystems, whose interaction supports economic development, yet leads to environmental damage, and the deterioration or improvement of the population's quality of life. In this manner, the continuous evaluation and forecasting of this behavior contributes to developing strategies to improve the habitability, viability and equity of urban territories with a view towards meeting the targets established by the SDGs.

While some studies have been developed to forecast sustainable development, these have focused either on specific sustainability dimensions, or on understanding countries' evolution regarding the same. The later are analyzed from a global perspective based on behavior in different territories. Along these lines, this study, which includes coordinating a series of procedures, contributes to the advancement of sustainability at the urban micro-territory scale. Its comprehensive method contributes to the academic and public arenas in the sense that it puts forth a tool that forecasts the category level of future sustainability in a micro-territory, such as Kennedy. It provides an opportunity to develop information gathering strategies and action plans, as well as monitor their implementation.

This instrument stands out in the sense that it reduces the territorial and temporal scope of information, in order to have a better territorial observation and to make use of systematized tools to analyze the portfolio of governmental proposals as techniques in different fields of sustainability, thus contributing to habitability, viability and equity interactions.

The micro-territory analyzed as a case study in this research study is representative of different environmental, social and economic conditions in Bogota. Kennedy is one of the most populated areas of the city, is one of the most polluted zones in Bogota in terms of air quality, in addition to having high levels of insecurity. It also represents an important economically active population of the city. The results from this study show consistent progress in implementing several policies and show the value of using statistical and machine learning tools to identify behavioral patterns of variables that influence the performance of micro-territories in the city, which is useful for decision-makers. Currently, decision-makers need to understand future situations regarding the implementation of current measures. Knowing of indicators that influence sustainable development enables leaders to make more informed decisions.

Concerning the results of the statistical analysis and the important variables through the Gini index in machine learning models, it is important to note that the later reinforces results from traditional methods.

This study found limitations on information availability for indicators that describe the behavior of sustainability dimensions in the micro territory. It is necessary to have a significant amount of information either for an appropriate characterization of each sustainability dimension, or to feed the machine learning models. Therefore, the information gathering phase required the most time and resources of this study.

Further research studies will be able to apply the methodology developed herein, in conjunction with machine learning models for each micro-territory in Bogota. The studies contemplate an analysis of micro-territories and how sustainable dimensions and their interactions are influenced by socio-economic aspects. This will enable a comparative analysis of the behavior of micro-territories, taking into account indicators on the environmental, social, and economic dimensions, as useful tools for decision-making related to resource prioritization and allocation. Additionally, conducting research that considers spatialized information will identify the behavior of habitability interactions and the viability of sustainable development in different territories.

Author Contributions: Conceptualization, Nidia Isabel Molina-Gómez; Data curation, Karen Rodríguez-Rojas and Dayam Calderón-Rivera; Formal analysis, Nidia Isabel Molina-Gómez, Karen Rodríguez-Rojas, Dayam Calderón-Rivera and Jose Luis Díaz-Arévalo; Investigation, Nidia Isabel Molina-Gómez; Methodology, Nidia Isabel Molina-Gómez; Software, Dayam Calderón-Rivera; Supervision, P. Amparo López-Jiménez; Visualization, Nidia Isabel Molina Gómez and Karen Rodríguez; Writing – original draft, Nidia Isabel Molina-Gómez; Writing – review & editing, Jose Luis Díaz-Arévalo and P. Amparo López-Jiménez. All authors have read and agreed to the published version of the manuscript.

Funding: "This research received no external funding".

Acknowledgments: The authors would like to thank the entities for the provision of information for this project development. Additionally, the authors are grateful for the support of professionals and the community for their contributions to the indicator's qualification.

Conflicts of Interest: "The authors declare no conflict of interest."

Supplementary Materials: Table S1: Annual indicators of Kennedy (2009-2017)

Indicator code	Sustainability Dimension	Interaction	Subject	Indicator	SDG	Year									Information source
						2009	2010	2011	2012	2013	2014	2015	2016	2017	
E1	Environmental	Livable	Air	Annual mean PM_{10} in Kennedy station ($\mu g/m^3$)	11	76.4	78.7	71.0	63.4	62.1	63.5	63.1	59.2	53.1	Bogota annual air quality report (2009-2017)
E2		Livable	Air	Annual mean ($PM_{2,5}$) in Kennedy station ($\mu g/m^3$)	11	25.4	35.1	28.4	25.9	26.7	29.5	28.8	27.6	24.3	
E3		Livable	Air	Exceedance of the national air quality standards for PM_{10} (%)	11	0.4	0.5	0.5	0.9	10.0	14.7	8.0	3.8	0.9	
E4		Livable	Air	Exceedance of the national air quality standards for $PM_{2.5}$ (%)	11	0.5	0.5	1.0	0.8	1.0	17.0	2.0	4.6	0.0	
E5		Livable	Air	Air quality index in Kennedy	11	91.9	92.9	84.9	102.0	76.0	83.0	74.0	82.0	80.0	

Indicator code	Sustainability Dimension	Interaction	Subject	Indicator	SDG	Year									Information source
						2009	2010	2011	2012	2013	2014	2015	2016	2017	
E6		Livable	Waste	Waste disposed of in landfills (Ton)	11	302,437.1	321,252.9	349,430.3	346,795.1	325,668.2	339,025.9	329,167.9	330,010.2	340,287.4	Official response from Special Administrative Unit for Public Services (UAESP in spanish). Document: 20188700044612 of 2019/02/18
E7		Livable	Waste	Construction and demolition waste with appropriate disposal (RCD) (Ton)	11	9,376.8	8,897.1	8,236.8	8,130.1	7,141.7	7,982.7	9,478.4	8,543.7	6,052.8	
E8		Livable	Water	Treated wastewater (%)	6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Official response from Acueducto de Bogotá. Document E-2018-023050 of 2018/03/20
E9		Livable	Water	Water quality of the river Fucha - section 3	6	35.0	31.0	37.0	31.0	41.0	29.0	36.0	48.0	27.0	Environmental observatory of Bogota. Distrital Secretariat of the Environment http://oab.ambientebogota.gov.co/indicadores/?id=433&v=l. Accessed on 2019/11/20
E10		Livable	Water	Water quality of the river Tunjuelo - section 4	6	35.0	41.0	44.0	38.0	43.0	45.0	40.0	44.0	53.0	
E11		Livable	Water	Water quality index	6	99.8	99.8	99.9	99.9	99.9	99.9	100.0	100.0	99.3	Distrital Secretariat of Planning. Portal geoestadístico. http://www.sdp.gov.co/gestion-estudios-estrategicos/informacion-cartografia-y-

														estadistica/portal-geoestadistico. Accessed on 2019.11.20
E12		Livable	Green space	Green spaces and recreation areas (km^2)	11	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	Distrital Institute of Recreation and Sports (IDRD in spanish). Available in https://www.idrd.gov.co/SIM/Parques/buscadorParques.php Accessed on 2019/11/10
E13		Livable		Number of trees per hectare	11	27.7	28.7	29.3	29.4	29.9	30.3	31.7	33.6	33.8	Official response from Botanical Garden of Bogota, José Celestino Mutis
E14		Sustainable		Protected areas (km^2)	15	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	Portal geoestadístico 2019.
S1	**Social**	Social	Health	Mortality rate of cardiopulmonary disease, pulmonary circulation diseases and other heart diseases	3	7.6	8.7	9.2	11.3	12.8	13.8	11.9	9.1	9.1	Bogota City Hall. Local diagnosis with social participation (2011, 2014, 2016). Atlas de la salud pública de la localidad de Kennedy 2016
S2		Social		Death rate due to chronic diseases	3	2.9	2.4	2.8	2.6	2.8	2.1	2.7	2.2	2.8	SALUDATA. Health observatory of Bogota. 2019

S3		Social	Health	Mortality rate of pneumonia in adults over 64	3	106.9	74.3	72.4	66.0	85.5	83.1	75.8	74.6	76.5	Official response from Distrital Secretariat of Health (SDS in spanish). Document 2019ER38259 of 2019/05/16
S4		Social		Mortality rate due to pneumonia in minors under 5	3	16.0	11.5	8.0	3.5	2.3	8.1	4.6	6.3	6.3	Bogota City Hall. Local diagnosis with social participation (2011, 2014, 2016). Atlas de la salud pública de la localidad de Kennedy 2016
S5		Social		Mortality rate due to severe acute respiratory infections for all ages	3	14.0	9.7	9.2	8.6	10.4	10.8	10.4	10.6	10.6	Official response from Distrital Secretariat of Health.
S6		Social		Infant mortality rate for all causes	3	9.9	11.3	10.7	11.6	10.2	8.9	9.5	9.3	9.1	Bogota City Hall. Local diagnosis with social participation (2011, 2014, 2016). Atlas de la salud pública de la localidad de Kennedy 2016
S7		Social		Acute malnutrition in children under 5 (proportion)	2	1.8	1.5	1.5	1.6	1.5	1.4	1.3	1.3	1.2	SALUDATA. Health observatory of Bogota. 2019

S8		Social		Chronic malnutrition in children under 5 (proportion)	2	16.6	15.9	14.6	15.6	14.6	18.9	14.8	15.2	15.6	
S9		Social		Global malnutrition in children under 5 (proportion)	2	2.3	3.5	0.0	0.0	0.0	1.2	1.2	0.0	0.0	
S10		Social		Mortality rate in children under 5 (proportion)	3	11.0	12.5	12.2	12.9	12.0	10.4	10.9	10.5	10.1	SALUDATA. Health observatory of Bogota. 2019
S11		Social	Demographic	Fertility rate	3	44.0	43.2	42.0	42.6	41.3	4.6	40.1	34.2	31.9	
S12		Social		Crude death rate	3	2.9	3.2	3.3	3.3	3.3	3.3	3.2	3.3	2.8	
S13		Social	Health	Bacterial meningitis rate	3	0.5	0.3	0.3	0.5	0.1	0.1	0.9	0.6	0.3	
S14		Social		Mortality rate in children under-5 years due to acute respiratory infection	3	4.6	1.2	5.7	11.5	2.3	3.5	5.7	4.2	5.3	
S15		Social		Dengue mortality rate	3	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	Official response from Distrital Secretariat of Health

S16		Social	Health	Mortality rate due to Diarrhoeal disease in children under-5	3	1.1	1.3	1.1	0.0	0.0	1.2	0.0	1.1	0.0	Atlas de la salud pública de la localidad de Kennedy (análisis de condiciones, calidad de vida, salud y enfermedad) 2017-2018
S17		Social		Infant mortality per 1,000 live births	3	159.0	178.0	164.0	181.0	155.0	134.0	143.0	131.0	121.0	Bogota City Hall. Local diagnosis with social participation (2011, 2014, 2016). Atlas de la salud pública de la localidad de Kennedy 2016
S18		Social		Perinatal mortality rate	3	9.9	11.3	10.7	11.1	10.3	8.9	9.5	9.3	9.1	
S19		Social		Maternal mortality ratio	3	17.7	22.7	25.4	23.2	13.4	12.3	12.4	13.9	13.8	SALUDATA. Health observatory of Bogota. 2019)
S20		Social		Hepatitis B rate per 100,000 population	3	3.8	4.8	4.8	4.0	3.2	4.4	3.3	3.8	0.7	Official response from Distrital Secretariat of Health.
S21		Social		Breastfeeding mortality rate	3	43.7	44.6	45.6	25.7	32.9	26.6	20.0	13.4	11.0	Bogota City Hall. Local diagnosis with social participation (2011, 2014, 2016). Atlas de la salud pública de la localidad de Kennedy 2016
S22		Equitable	Education	Mean years of schooling	4	8.3	8.5	8.8	9.1	9.4	9.6	9.9	10.4	10.7	Official response from Distrital Secretariat of Education (SDE

S23		Equitable		School attendance rate	4	81.5	81.8	84.7	85.5	86.7	88.1	90.0	91.5	93.5	in spanish). Document E-2109-82955 of 2019/05/15
S24		Equitable		Crude education coverage rate	4	85.4	85.4	90.8	89.7	86.8	87.7	84.2	72.1	74.0	
S25		Equitable		Dropout rate	4	2.8	2.9	2.7	2.9	2.4	1.9	2.3	2.7	2.7	
S26		Equitable		Population with primary and secondary education	4	107,751.0	139,339.0	173,530.1	203,545.0	202,186.0	200,994.0	200,086.0	222,563.0	251,374.1	
S27		Equitable		Illiteracy rate	4	1.6	1.6	1.6	1.4	1.4	1.2	1.2	1.2	1.1	
S28		Sustainable	Demographic	Life expectancy at birth (years)		76.0	76.0	77.1	77.1	77.1	77.1	77.1	77.1	77.1	Bogota City Hall. Local diagnosis with social participation (2011, 2014, 2016, 2017)
S29		Social		Birth rate	3	16.1	15.6	15.0	15.2	14.6	14.3	14.0	11.9	11.0	SALUDATA. Health observatory of Bogota. 2019
S30		Sustainable		Population ageing rate		8.4	8.4	11.0	9.8	10.0	9.6	10.1	11.9	11.9	Bogota City Hall. Local diagnosis with social participation (2014, 2016, 2017). Atlas de la salud pública

S31		Equitable	Security	Number of firearm deaths, homicides, fights and/or clashes	16	127.0	126.0	183.0	124.0	101.0	90.0	84.0	84.0	60.0	Official response from Metropolitan Police of Bogota. Document 1130652019 of 2019/05/16
S32		Equitable		Reports of domestic, family and child violence and abuse (number)	5	369.0	398.0	410.0	270.0	258.0	912.0	1,760.0	3,409.0	3,491.0	
S33		Equitable		Simple and aggravated theft (number)	16	1,779.0	1,549.0	1,587.0	2,929.0	2,573.0	2,860.0	4,121.0	4,291.0	6,651.0	
S34		Equitable		Vehicle theft (number)	16	455.0	435.0	460.0	508.0	416.0	505.0	693.0	697.0	616.0	
S35		Sustainable	Transport	Passengers transported by the mass public transport system (number)	9	36,437,561.0	41,731,722.0	45,236,883.0	46,734,903.0	48,455,774.0	51,038,712.0	50,570,476.0	61,366,636.0	61,175,936.0	Official response from TransMilenio. Document: 2019-ER-15605 of 2019/05/15
S36		Sustainable	Demographic	Population density per km^2	11	2.6	2.9	2.8	2.9	3.0	3.2	3.2	3.3	3.4	Official response from Distrital Secretariat of Planning (SDP in spanish): Document 2-2019-39192 of 2019/06/17

S37		Social		Population located in flood-prone areas	11	1,030,903.0	1,052,725.0	1,075,024.0	1,097,601.0	1,120,274.0	1,142,901.0	1,165,318.0	1,187,315.0	1,208,980.0	Official response from Distrital Institute for Risk Management and Climate Change (IDIGER in spanish). Document 2018ER16615 of 2019/09/27
S38		Livable		Flood-prone areas (km^2)	11	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	38.5	
S39		Livable	Transport	**Number of road traffic deaths**	3	2.0	6.0	2.0	7.0	3.0	8.0	2.0	3.0	1.0	Official response from Metropolitan Police of Bogota. Document 1130652019 of 2019/05/16
S40		Social	Health	Suicide (number)	3	19.0	26.0	31.0	21.0	30.0	32.0	28.0	47.0	29.0	
S41		Livable	Demographic	Informal settlements (km^2)	11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Official response of Distrital Secretariat of Planning.
S42		Equitable	Service coverage	Households with access to natural gas service (number)	1	25,777.0	35,586.0	45,381.0	53,811.0	61,835.0	68,122.0	76,629.0	84,358.0	90,390.0	Official response from Gas Natural Fenosa. Document 02410516 of 2019/05/30
S43		Equitable		Population with access to natural gas service at home	1	937,698.0	946,765.0	956,712.0	967,879.0	977,471.0	966,724.0	1,143,177.0	1,132,699.0	1,186,010.0	
S44		Equitable		Coverage of storm water drainage system (%)	1;6	94.4	96.9	97.7	97.9	99.3	99.7	99.5	99.2	99.4	Official response from Acueducto de Bogotá. Document E-

S45		Equitable		Proportion of population using safely managed drinking-water services	1;6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	2018-023050 of 2018/03/20
S46		Equitable		Population with sewerage services (%)	1;6	100.0	100.0	100.0	100.0	99.9	99.9	99.9	99.9	100.0	
S47		Equitable	Service coverage	Access to health services (population)	1;3	2,556,316.0	2,935,223.0	3,032,134.0	3,125,321.0	3,226,789.0	3,309,321.0	3,393,622.0	3,490,449.0	3,567,041.0	Bogota City Hall. Local diagnosis with social participation (2014, 2016, 2017). Atlas de la salud pública
EC1	**Economic**	Economic	Income and expenditures	Gross domestic product per capita	8	0.4	0.4	0.5	0.4	0.6	0.5	0.7	0.4	0.3	Bogota City Hall. Local diagnosis with social participation (2014, 2016, 2017). Official Response from National Administrative Department of Statistics (DANE): 20191210166272 from 2019.08.27
EC2		Equitable	Employment	Unemployment rate	8	7.5	7.5	7.3	7.4	7.4	7.5	7.4	7.4	7.4	Bogota City Hall. Local diagnosis with social

EC3		Economic		Employed population aged 12-64	8	432,001.0	479,210.0	499,319.0	500,098.0	505,610.0	507,587.0	540,680.0	566,760.0	571,821.0	participation (2014, 2016, 2017)
EC4		Economic		Economically active population		483,082.0	501,067.0	538,905.0	543,787.0	542,467.0	526,798.0	575,678.0	609,543.0	620,053.0	
EC5		Economic	Income and expenditures	Per capita household income (COP)	8	567,895.0	623,567.0	636,033.0	743,560.0	801,589.0	964,398.0	970,678.0	975,456.0	981,690.0	Official response of SDP. Document 2-2019-39192 of 2019/06/17
EC6		Economic	Economic structure	Affiliation of assurance to the general health safety system (%)	1;3	85.0	87.0	87.0	88.0	89.0	91.0	91.0	90.0	89.0	SALUDATA. Health observatory of Bogota. 2019
EC7		Equitable		Access to electricity (population)	1;7	999,693.0	999,693.0	1,016,545.5	1,033,111.9	1,049,456.4	1,066,087.9	1,082,852.3	1,074,626.0	1,208,980.0	Official response from Enel-Codensa. Document 07557011 of 2019/06/17
EC8		Equitable	Poverty	Population below poverty line	1	53,875.0	135,653.0	135,653.0	53,146.0	45,851.0	114,909.0	183,966.0	167,755.0	151,544.0	Bogota City Hall. Local diagnosis with social participation (2009, 2011, 2014)
EC9		Equitable		Unsatisfied basic needs	1	5.4	5.1	5.1	5.1	4.4	4.7	4.9	9.5	14.2	
EC10		Equitable		Population at high risk due to water shortage	1	0.0	0.0	0.0	151.0	233.0	130.0	149.0	223.0	142.0	Official response from Acueducto de Bogotá. Document E-2018-023050 of 2018/03/20

EC11		Economic	Income and expenditures	Taxes on movable or immovable property (COP)		55,500,820.0	67,518,745.0	85,939,359.0	100,226,926.0	115,060,196.0	131,155,566.0	161,264,822.0	172,735,979.0	182,989,654.0	Official response of Distrital Secretariat of Treasury. Document 2019EE168841 of 2019/09/12
EC12		Economic	Consumption and production	Energy consumption (COP)	7	182,177,891,674	201,435,901,662	226,065,938,744	234,705,238,066	237,409,670,668	261,695,983,279	283,207,051,836	315,296,562,822	320,879,325,401	Official response from Enel-Codensa. Document 07557011 of 2019/06/17
EC13		Viable		Energy consumption (kWh/year)	7	614,859,738	639,174,551	656,614,603	670,452,228	668,440,731	702,897,217	740,591,385	733,781,327	734,076,230	
EC14		Economic	Income and expenditures	Gini coefficient	1	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	Bogota City Hall. Local diagnosis with social participation (2009, 2011, 2014, 2017)
EC15		Viable	Consumption and production (Transport)	Arterial roads in good conditions (km)	11	157.3	174.1	108.6	143.7	158.0	153.0	196.3	378.3	398.4	Portal geoestadístico 2019.
EC16		Sustainable		Per capita water consumption (residential) (m^3/hab-d)		0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	Official response from Acueducto de Bogotá. Document E-2018-023050 of 2018/03/20
I1	**Institutional**	Sustainable	Government	Sustainable development policies or strategies (number of policies)	17	5.0	4.0	5.0	5.0	7.0	6.0	5.0	5.0	5.0	Bogota City Hall. Local diagnosis with social participation (2009, 2011, 2017). Government plans of Bogota 2009, 2011, 2016

I2		Equitable		Public spending on education (COP)	1	216,624,778,909.2	222,594,202,649.0	229,738,286,481.8	241,721,053,673.8	237,029,077,484.1	275,722,736,431.9	287,451,899,003.8	318,012,306,088.9	391,475,212,039.6	Official response from Distrital Secretariat of Education. Document E-2109-82955 of 2019/05/15
I3		Social		Gross participation rate		48.4	55.5	61.1	57.6	61.1	60.0	61.1	62.1	60.8	Bogota City Hall. Local diagnosis with social participation (2009, 2011, 2017)
I4		Social	Community and social services	Cultural events (number of events)		2,008.0	2,060.0	2,116.0	2,172.0	2,228.0	2,284.0	2,340.0	2,364.0	2,480.0	

Sustainable Development Goals (SDG)
1 No poverty
2 Zero hunger
3 Good health and well-being
4 Quality education
5 Gender equality
6 Clean water and sanitation
7 Affordable and clean energy
8 Decent work and economic growth
9 Industry, innovation and infrastructure
10 Reduced inequalities
11 Sustainable cities and communities
12 Responsible consumption and production
13 Climate action
14 Life below water
15 Life on land
16 Peace, justice and strong institutions
17 Partnerships for the goals
* Indicators in bold type were also included in the machine learning models fed with monthly information

References

1. United Nations, Department of Economic and Social Affairs. World urbanization prospects The 2018 Revision. New York, 2019.
2. Shen, L.; Kyllo, J.; Guo, X. An Integrated Model Based on a Hierarchical Indices System for Monitoring and Evaluating Urban Sustainability. *Sustainability*. **2013**, 5, 524–559.
3. Verma, P.; Raghubanshi, A.S. Urban sustainability indicators: Challenges and opportunities. *Ecol. Indic.* **2018**, 93, 282–291.
4. Phillis, Y.A.; Kouikoglou, V.S.; Verdugo, C. Urban sustainability assessment and ranking of cities. *Comput. Environ. Urban Syst.*, **2017**, 64, 254-265.
5. WCED, Report of the World Commission on Environment and Development: Our Common Future: Report of the World Commission on Environment and Development. Oslo, 1987.
6. Gerry Marten, Human Ecology: Basic Concepts for Sustainable Development - Populations and Feedback Systems. Available online: http://gerrymarten.com/ecologia-humana/capitulo02.html (accessed: 28-Jan-2020).
7. Tanguay, G.A.; Rajaonson, J.; Lanoie, P. Measuring the sustainability of cities : An analysis of the use of local indicators. *Ecol. Indic.* **2010,** 10, 407–418.
8. Carrillo-Rodríguez, J.; Toca, C.E. Sustainable performance in Bogota: Building an indicator from local performance. *Eure*, **2013,** 39, 165–190.
9. Mapar, M.; Jafari, M.J.; Mansouri, N.; Arjmandi, R.; Azizinejad, R.; Ramos, B. Sustainability indicators for municipalities of megacities: Integrating health, safety and environmental performance. *Ecol. Indic.* **2017,** 83, 271–291.
10. Rajaonson, J.; Tanguay, G.A. A sensitivity analysis to methodological variation in indicator-based urban sustainability assessment: a Quebec case study. *Ecol. Indic.*, **2017,** 83, 122–131.
11. Toumi, O.; Le Gallo, J.; Ben Rejeb, J. Assessment of Latin American sustainability. *Renew. Sustain. Energy Rev.* **2017,** 78, 878–885.
12. Li, Y.; Wu, Y.-X.; Zeng, Z.-X.; Guo, L. Research on forecast model for sustainable development of Economy-Environment system based on PCA and SVM. in *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, 2006; pp. 3590–3593.
13. Zhang, Y.; Huan, Q. Research on the evaluation of sustainable development in Cangzhou city based on neural-network-AHP, in *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, 2006, 3144–3147.
14. Pérez-Ortíz, M.; de La Paz-Marín, M.; Gutiérrez, P. A.; Hervás-Martínez, C. Classification of EU countries' progress towards sustainable development based on ordinal regression techniques. *Knowledge-Based Syst.* **2014,** 66, 178–189.
15. Zhang, Y.; Shang, W.; Wu,Y. Research on sustainable development based on neural network. in *2009 Chinese Control and Decision Conference*, 2009, 3273–3276.
16. Dizdaroglu, D. Developing micro-level urban ecosystem indicators for sustainability assessment. *Environ. Impact Assess. Rev.* **2015,** 54, 119–124.
17. Distrital Secretariat of Planning, Monograph 2017 Diagnosis of the main territorial, infrastructure, demographic and socio-economic aspects Kennedy Locality 08. Bogotá, Colombia, 2018.
18. United Nations. Global indicator framework for Sustainable Development Goals and Agenda 2030 for Sustainable Development. 2018.
19. United Nations. Indicators of Sustainable Development : Guidelines and Methodologies. 3th ed., no. October. New York: United Nations, 2007.
20. Niemeijer, D.; de Groot, R. S. A conceptual framework for selecting environmental indicator sets. *Ecol. Indic.* **2008**, 8, 14–25.
21. Quiroga Martínez, R.; Stockins, P.; Holloway, M; Taboulchanas, K.; Sanchez, A. Methodological guide for developing environmental and sustainable development indicators in Latin American and Caribbean

countries. Santiago de Chile: CEPAL. United Nations. Economic Commission for Latin America and the Caribbean., 2009.
22. Scipioni,A.; Mazzi, A.; Mason, M.; Manzardo, A. The Dashboard of Sustainability to measure the local urban sustainable development : The case study of Padua Municipality. *Ecol Ind.* **2009,** 9, 364–380.
23. Alpopi, C.; Manole, C.; Colesca, S.E. Assessment of the sustainable urban development level through the use of indicators of sustainability. *Theor. Empir. Res. Urban Manag.* **2011,** 6, 78–87.
24. Cecchini, S. *Indicadores sociales en América Latina y el Caribe*. Santiago de Chile: Naciones Unidas, CEPAL, División de Estadísticas y Proyecciones Económicas, 2005.
25. Klopp, J.M; Petretta, D.L. The urban sustainable development goal: Indicators, complexity and the politics of measuring cities. *Cities*, **2017**, 63, 92–97.
26. Shen, Y.J.; Ochoa, J.; Shah, M.N.; Zhang, X. The application of urban sustainability indicators - A comparison between various practices. *Habitat Int.* **2011**, 35, 17-29.
27. Hák, T.; Janoušková, S.; Moldan, B. Sustainable Development Goals: A need for relevant indicators. *Ecol. Indic.*, **2016,** 60, 565–573.
28. Escobar, L. Synthetic indicators of environmental quality: a general model for large urban areas. *Rev. eure*, **2006**. XXXII, 73–98.
29. Sotelo, J. A.; Tolón, A.; Lastra, X. Indicators for and by sustainable development, a case study. *Estud. Geográficos*. **2012,** 72, 611–654.
30. Dizdaroglu, D. Developing micro-level urban ecosystem indicators for sustainability assessment. *Environ. Impact Assess. Rev.* **2015**. 54, 119–124.
31. Feleki, E.; Vlachokostas, C.; Moussiopoulos, N. Characterisation of sustainability in urban areas: An analysis of assessment tools with emphasis on European cities. *Sustain. Cities Soc.* **2018**. 43, 563–577.
32. Ocampo, L.; Ebisa, J.A.; Ombe, J.; Geen Escoto, M. Sustainable ecotourism indicators with fuzzy Delphi method – A Philippine perspective. *Ecol. Indic.* **2018,** 93, 874–888.
33. Distrital Secretariat of Planning, "Knowing Kennedy: Diagnosis of physical, demographic and socioeconomic aspects," Bogota, 2009.
34. SALUDATA- Health observatory of Bogota 2019. Available online: http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/salud-ambiental/consultaurgencias14anios/ (accessed: 28-Feb-2019).
35. Hospital del Sur. E.S.E. Local diagnosis with social participation 2014 locality of Kennedy. Bogotá, 2014.
36. Distrital Secretariat of the Environment. Bogota annual air quality report 2009. Bogota, Colombia, 2010.
37. Distrital Secretariat of the Environment. Bogota annual air quality report 2012. Bogota, Colombia, 2013.
38. Distrital Secretariat of the Environment. Bogota annual air quality report 2015. Bogota, Colombia, 2016.
39. Local Mayor of Kennedy. Local Risk Management and Climate Change Council General Characterization of Risk Scenarios. Bogota, 2018.
40. Distrital Secretariat of the Environment. Bogota annual air quality report 2014. Bogota, Colombia, 2015.
41. Distrital Secretariat of the Environment. Bogota annual air quality report 2013. Bogota, Colombia, 2014.
42. Distrital Secretariat of the Environment. Bogota annual air quality report 2011. Bogota, Colombia, 2012.
43. Distrital Secretariat of the Environment. Bogota annual air quality report 2010. Bogota, Colombia, 2011.
44. Portal geoestadistico 2019. Available online: http://www.sdp.gov.co/gestion-estudios-estrategicos/informacion-cartografia-y-estadistica/portal-geoestadistico (accessed: 20-Nov-2019).
45. Mayor of Bogota. Local diagnosis with social participation 2009-2010 locality of Kennedy. Bogota, Colombia, 2010.
46. Mayor of Bogota. Local environmental plan Kennedy better for all locality example for all 2017-2020. 2016.
47. Torres-Delgado A.; López Palomeque, F. The ISOST index: A tool for studying sustainable tourism. *J. Destin. Mark. Manag.* **2018,** 8, 281–289.
48. Cui, X.; Fang, C.; Liu, H.; Liu, X. Assessing sustainability of urbanization by a coordinated development index for an Urbanization-Resources-Environment complex system: A case study of Jing-Jin-Ji region, China. *Ecol. Indic.* **2019,** 96, 383–391.

49. Saaty, R. W. The analytic hierarchy process-what it is and how it is used. *Math. Model.* **1987**, 9, 161–176.
50. Schuschny, A.; Soto, H. "Guía metodológica Diseño de indicadores compuestos de desarrollo sostenible Andrés Schuschny," *Cepal*, p. 109, 2009.
51. Rokach, L.; Maimon, O. *Data mining with decision trees : theory and applications*, 2nd ed. Singapore: World Scientific Publishing Co. Pte. Ltd. 5, 2015.
52. Kuhn, M. *Applied Predictive Modeling in R*. 2014 .
53. Cortes, C. and Vapnik,V. Support-Vector Networks. *Mach. Learn.*, **1995**, 20, 273–297.
54. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T. Package ' caret .' 2013.
55. Ripley, B.; Venables, W. Feed-Forward Neural Networks and Multinomial Log-Linear Models. *February*. 2016.
56. Meyer, D.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C.-C.; Lin, C.-C. Package 'e1071.' 2019.
57. Shen, L.-Y.; Ochoa, J. Jorge.; Shah, M.N.; Zhang, X. The application of urban sustainability indicators – A comparison between various practices. *Habitat Int.* **2011,** 35, 17–29.

Apéndice C.

Air quality and urban sustainable development: the application of machine learning tools

Documento versión del autor publicado en la Revista: International Journal of Environmental Science and Technology (Electronic ISSN 1735-2630; Print ISSN 1735-1472). Recibido el 10 de Abril de 2020; revisado el 27 de Junio de 2020; aceptado el 11 de Agosto de 2020; publicado el 19 de Agosto de 2020. DOI: https://doi.org/10.1007/s13762-020-02896-6

Nidia Isabel Molina-Gómez[1,2], José Luis DíazArévalo[3] , P. Amparo López-Jiménez[2]

[1.]Department of Environmental Engineering, Universidad Santo Tomás, Bogotá, Colombia

[2.]Hydraulic and Environmental Engineering Department, Universitat Politècnica de València, Valencia, Spain

[3.]Department of Civil and Agricultural Engineering, Universidad Nacional de Colombia, Bogotá, Colombia

---

El análisis de diversas herramientas de aprendizaje automático, aplicadas al pronóstico de la calidad del aire y del desarrollo sostenible, sentó las bases para la elección de los métodos, métricas de desempeño y procesos clave para la aplicación de aprendizaje automático en la clasificación de los niveles de desarrollo sostenible, en la relación de calidad del aire y salud y en la definición de las variables de mayor influencia.

ABSTRACT

Air quality has an effect on a population's quality of life. As a dimension of sustainable urban development, governments have been concerned about this indicator. This is reflected in the references consulted that have demonstrated progress in forecasting pollution events to issue early warnings using conventional tools which, as a result of the new era of big data, are becoming obsolete. There is a limited number of studies with applications of machine learning tools to characterize and forecast behavior of the environmental, social, and economic dimensions of sustainable development as they pertain to air quality. This article presents an analysis of studies that developed machine learning models to forecast sustainable development and air quality. Additionally, this paper sets out to present research that studied the relationship between air quality and urban sustainable development to identify the reliability and possible applications in different urban contexts of these machine learning tools. To that end, a systematic review was carried out, revealing that machine learning tools have been primarily used for clustering and classifying variables and indicators according to the problem analyzed, while tools such as artificial neural networks and support vector machines are the most widely used to predict different types of events. The non-linear nature and synergy of the dimensions of sustainable development are of great interest for the application of machine learning tools.

Key words: air pollution, sustainability, forecasting, sustainable development goals, and influencing variables.

1 INTRODUCTION

Compatible, socially just, and economically viable ecological development is at the heart of the concept of sustainable development (Mellos, 1988). This term has been more widely recognized since the Earth Summit, where it was established that it is essential for development to meet the needs of the present without compromising the ability of future generations to meet their own needs (WCED, 1987). Sustainable development is a goal of the global agenda, which nations have pursued since the Millennium Development Goals and currently with the Sustainable Development Goals (SDGs). These set targets that should be integrated into countries' national development plans. In this regard, considering that it is of special concern for nations to achieve the SDGs, there is great interest in forecasting SD behavior. Nations need to make progress in forecasting their territorial behavior and consider indicators' variability in the scope of sustainable development. A future vision of this behavior for the cities and regions that comprise nations is a valuable tool for decision-makers.

Sustainable development (SD henceforth) is supported by three encompassing dimensions; the environmental, social, and economic dimensions. This concept includes environmental health, social equality and economic growth as part of the dimensions' interactions, which on balance, are difficult to measure (Lubell et al., 2009). Any change of any one of those interactions could impact SD, demonstrating that balancing and integrating these dimensions is challenging in the quest for sustainability. In this sense, it is important to mention that greater than 55% of the world population lives in urban zones (United Nations 2019). As a result, air, water and soil quality, species extinctions, worsening health of populations, poverty, among others, all impact SD, specifically urban sustainable development.

Air pollution is one of the effects of great concern at the global level. It is included in the SDGs and may become the main environmental cause of premature mortality (Cruz et al., 2017). A study performed by Krzyzanowski et al. (2014) identified that at least 96% of the population of large cities is exposed to particulate matter smaller than 2.5 micron ($PM_{2.5}$). This study also identified that cities with higher concentrations of particulate matter and the lowest air quality improvement rates over the last decade, tend to be countries with lower levels of economic development. Furthermore, 92% of the world's population lives in places where the air quality exceeds standards established by the World Health Organization (WHO) (WHO 2016). Moreover, the annual global mortality of nearly 3 million people is related to exposure to outdoor air pollution (WHO, 2016). Instruments have been developed to monitor and forecast atmospheric pollutants in order to control these correlations.

Several studies have applied machine learning (ML) tools to forecast air quality and certain pollutants, primarily particulate matter smaller than 10 and 2.5 micrometers (PM10 and PM2.5). For example, Antanasijević et al. (2013) and Paas et al. (2017) applied artificial neural networks (ANN), Karimian et al. (2019) and Zhou et al. (2019) used deep neural networks (DNN), and Oprea et al. (2016) and Wang (2019) utilized decision trees (DT) or their ensembles. Furthermore, support vector machine (SVM) models were developed for the spatio-temporal predictions of PM2.5 (Song et al. 2014; de Hoogh et al. 2018).

Machine learning tools are instruments that forecast the behavior of different variables considering large volumes of data and in many cases, substantial quantities of predictors. With several advantages over conventional models, the use of ML has advanced. Different documents on data mining and ML extensively describe the algorithms and their applications (Brink et al. 2016; Lässig et al. 2016). Those studies, which applied conventional models and ML tools, analyzed and forecasted the behavior of pollutants for certain regions and territories. The principal aim has been to provide useful technological tools and results for decision-makers in order to protect populations' health. Nevertheless, these studies have not identified air pollution's impacts on SD. Learning the influence of air pollution on nations' sustainability, with respect to mortality rates and the impacts of air pollution on the world population is of great interest.

Machine learning tools are useful for understanding the impacts of air pollution on the sustainable development of territories. As such, the primary objective of this paper is to analyze the work developed in prior studies, which have used ML tools to forecast air quality and SD. This study aims to identify tools that can closely relate both concepts in order to identify the state of the art ML applications, existing gaps in the field, and determine

aspects that can be addressed in the future. This research paper develops its analysis on the following questions: which machine learning tools have been applied to forecast the sustainable development behavior of territories, and which machine learning tools could be applied to identify the influence of air pollution on sustainable development?

This study is of special interest for governments, scientific communities, and societies. It is innovative in that it provides tools for environmental governance. Sustainable development is a challenge for governments as it requires an understanding of each component's behavior, the sustainability dimensions and the territories' evolution in terms of the SDGs. Additionally, this study provides specific information for decision-makers by analyzing the application of machine learning tools where technology and development issues meet.

This paper is structured into four sections; following this introduction, the second section Materials and Methods, describes the procedures applied in the systematic revision. The next section introduces the results through a description of the different ML tools employed in various studies to forecast air quality and SD levels. This section contains the results and subsequent discussion of this review, evaluating the tools' functionality based on the requirements for their application in the integration of both concepts. This is followed by a section that analyzes the gaps and proposed future developments. The paper closes with the conclusions section.

2 MATERIALS AND METHODS

This study was developed based on the following four relevant stages: 1) determining the research question and establishing the criteria for exhaustive literature selection; 2) preparing a dataset with the research documents collected in stage 1; 3) researching and evaluating records, and 4) analyzing the selected literature. Figure 1 presents the methodological framework used for this research study.

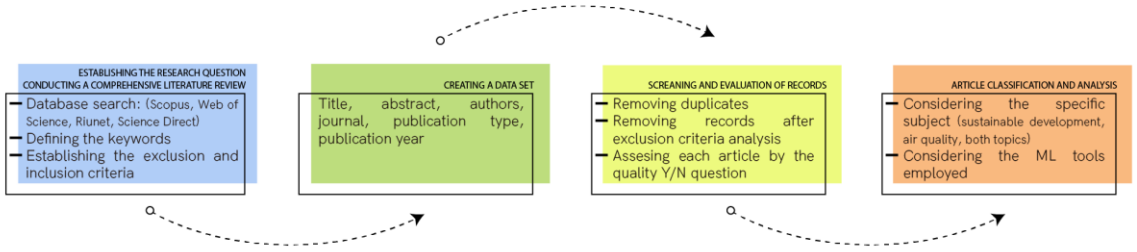


Fig. 1 Methodological framework to understand the relationship between air quality and sustainable development for forecasting.

2.1 ESTABLISHING THE RESEARCH QUESTION AND CONDUCTING A COMPREHENSIVE LITERATURE REVIEW

Sustainable urban development encompasses the integration of three fundamental pillars and their interactions. Identifying the machine learning tools used to forecast sustainable development in territories will enable the identification of a set of tools that could be implemented in other territories with similar characteristics. Furthermore, this identification would be able to guide the combination of ML methods to solve specific problems in each SD dimension, particularly related to air pollution. Therefore, to answer the research question, the criteria for searching and choosing scientific literature have been established as is described in Table 1.

Table 1. Primary aspects to identify scientific literature in the framework of this study

Database	Keywords and equation used in the search	Inclusion and exclusion criteria
- Scopus - Web of Science (WoS) - Riunet	((("machine learning" OR "data mining" OR "pattern recognition") AND ("air quality" OR "air pollution" OR "atmospheric pollution" OR "atmosphere" OR "atmosp*") AND ("sustainable development" OR "sustainability" OR "sustainable" OR "sustain*")))	**Date:** Last 20 years (2000-2019) **Publication type:** Articles, reviews, doctoral theses, conference papers. **Language:** English **Peer reviewed publications:** Articles, reviews, doctoral theses, conference papers published in a peer reviewed journals or conference proceedings. **Publications' objective:** The primary goal of the publication is forecasting sustainable development OR air quality OR atmospheric pollution with machine learning tools. Only studies that strictly entailed the use of machine learning tools in both sustainable development and air quality were included.
Science Direct	((("machine learning" OR "data mining") AND ("air quality" OR "air pollution" OR "atmospheric pollution" OR atmosp) AND ("sustainable development" OR sustain)))	

The databases identified in Table 1 were the primary sources for data collection. Additionally, the search engine of the Universitat Politècnica de València was used based on the criteria outlined in Table 1.

2.2 CREATING A DATA SET FOR THE RECORDED INFORMATION

A template was designed for the inclusion of each of the records gathered by the searches. In addition to specific information identifying each record, the information base included the identification of the machine learning tools used in the studies, their use in each case, the tools utilized in processes to train and validate the data for forecasting, and each study's analysis objective: sustainable development, air quality according to pollutants or sustainable development dimensions, in the event that the study focused on specific dimensions.

2.3 SCREENING AND EVALUATION OF RECORDS

Once the data set was created based on the criteria established in Table 1, the identified records were reviewed. The information was verified by reading abstracts of each study. Duplicate records in the databases were eliminated, as well as records that did not meet the quality criteria established for inclusion in the data set to be analyzed.

2.4 ARTICLE CLASSIFICATION AND ANALYSIS

The set of records was then analyzed by reading each document and identifying essential information to answer the following research questions: which machine learning tools have been applied to forecast the sustainable development behavior of territories, and which machine learning tools could be applied to identify air pollution's impact on sustainable development?

The classification and analysis of the scientific literature were guided by the subjects covered in the documents, the machine learning tools applied, how they were used, the evaluation metrics employed, and their performance.

3 RESULTS AND DISCUSSION

Following the methodological framework described above, a great number of scientific works were found, mainly in the fields of forecasting air pollutants, air quality indexes to issue early warnings, ensembles of tools to identify future spatio-temporal behavior of pollutants, as well as the identification of factors that induce the presence of atmospheric pollutants at a specific level. Although scientific production has increased with the integration of sustainable development goals, research on the application of machine learning tools in this field is limited. The following are the results found in each case.

3.1 APPLICATION OF MACHINE LEARNING TOOLS IN THE FRAMEWORK OF SUSTAINABLE DEVELOPMENT

Different studies have evaluated the state of progress of the environmental, social, and economic dimensions of SD and their respective indicators. Machine learning tools have analyzed the behavioral patterns of conjugate variables, as well as the lack of a self-learning capacity, the treatment of non-linear relationships, and possible bias to estimate the weight of each dimensions of sustainable development (Zhang et al., 2009; Zhang and Huan, 2006).

Machine learning encompasses tools and techniques to identify patterns within data or process time series in some cases. ML is an integral piece of predictive modeling; whose principal aim is developing reliable models to forecast variable behavior or that of specific problems. A model's reliability and accuracy are determined by specific metrics whose application is closely tied the model's purpose. The root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), maximum mean absolute error (MMAE), mean percentage error (MPE), mean absolute percentage error (MAPE), mean bias error (MBE), accuracy (Acc), relative error ratio (RER), correlation coefficient (R), and index of agreement (IA) are among the metrics used to evaluate the performance of ML models. ML has been successfully applied to specific subjects regarding the dimensions of SD. Certain studies have examined the dimensions and specific indicators to a greater degree (Gounaridis et al., 2018; Li et al., 2006a; Madu et al., 2017), while others employ a more holistic vision (Zhang et al., 2009; Zhang and Huan, 2006). Studies that assessed environmental, social and economic policies are included, and their results provide an understanding of the variables to be reinforced to achieve certain SDGs and increase eco-efficiency and socio-efficiency levels in the territory analyzed. The application of ML tools depends on the objectives established and information availability.

3.1.1 Information required for the studies

The application of ML tools both in the scope of the indicator and in a broader vision of SD, requires information from macroeconomic indicators, such as gross domestic product (GDP), level of financial development, education, rate of employment generated by industry, and environmental quality indexes (Li et al., 2006; Pérez-Ortíz et al., 2014; Zhang et al., 2009). This information is centered on indexes, indicators, and variables according to the dimension analyzed. To illustrate, for the environmental dimension; the concentration of atmospheric pollutants, water consumption and volume of waste-water per unit of GDP, the biochemical oxygen demand, noise level, the rate of treated industrial water, solid waste use, and reforested areas are taken into consideration (Li et al. 2006; Zhang and Huan 2006). With respect to the social dimension; the student rate, population density, unemployment rate, distance to hospitals and transportation stations are used (Zhang and Huan 2006; Gounaridis et al. 2018). To analyze the economic dimension; the GDP, the ratio of industry to GDP, foreign investment, energy consumption per unit of GDP, among others, are taken into account (Zhang and Huan 2006; Wang and Xiao 2017). However, this data is not available in every territory with the necessary quality and timeliness, and it is generally recorded at the national level. Most studies in this field analyze the behavior of cities and provinces in China, others performed behavioral analysis of different European Union nations, primarily applying SVM, with ANNs and DTs employed to a lesser degree.

SD assessments are limited due to the scarcity and quality of data (Toumi et al., 2017), as it tends to present atypical values, additional noise, missing data and even errors (Gibert et al., 2016). Countries in Latin America and the Caribbean have deficiencies in the standardized information management and collection. Some cities in the region are characterized by high population densities, disjointed urban development, and interesting geographic conditions for the study of air quality and sustainable urban development. Forecasting the behavior in a territory as it pertains to SD in these types of urban areas calls for the application of tools to understand their growth and development, as well as to complete and standardize the information to be included in a ML model.

The scarcity of information on the characteristics required to calculate and forecast sustainability progress, demonstrates the need to use tools to assign data to the evaluated territories, as presented in the study developed by Phillis et al. (2017). Other studies have proposed the spatial identification of different variables (Holloway and Mengersen, 2018), but they are conditioned upon the information visible in satellite images or georeferenced data. These studies provide useful tools to solve the problem of data quality and territorial scope.

Nevertheless, it is important to solve the problem of data frequency generation to feed and validate machine learning models. In this sense, intervention from, and synergy with, the government sector plays an important role.

Data capturing, recording and quality assurance is another challenge that impacts SD as it limits the effectiveness and accuracy of ML application results. Territories have centralized their information through reports or international management platforms. There are different types of software and programs that provide support for informed decision-making (Gibert et al., 2012; Kadiyala and Kumar, 2017a), but a minimum quality standard must be ensured for the information used.

3.1.2 Machine learning models

A selected group of machine learning tools is described below. This group of tools focuses on classification and regression problems in the framework of sustainability analysis and air quality forecasting. Through an analysis of different studies developed in the last 20 years related to forecasting sustainable development, it was found that SVMs, ANNs, Random Forest (RF), and DNNs are the primary machine learning tools used to forecast sustainable development or their dimensions in different territories.

- Support Vector Machine (SVM)

SVMs are a group of supervised learning algorithms related to classification and regression problems (Sierra, 2006). The classification categorizes new objects in two or more separate groups based on their properties and a set of observations (de Hoogh et al., 2018). SVM draws a vector to separate classes; the greater their separation, the greater the recognition of different groups. For problems regarding more than two dimensions, SMV finds a hyperplane that maximizes the separation margin between classes. The classification could be supported by Kernel functions. The regression, referred to as support vector regression (SVR), enables it to be applied to forecast continuous variables; as such, the data needs to be trained (Suárez et al., 2011). Both classification and regression require data for validation and testing purposes.

SVMs have advantages in treating small samples and non-linear data patterns (de Hoogh et al., 2018; Wang and Xiao, 2017). They are used to predict nations' progress levels in terms of the environmental, social, economic, and institutional dimensions of SD, and to forecast eco-efficiency, including the analysis of spatial data behavior and mixed frequency data modeling (Li et al., 2006; Pérez-Ortíz et al., 2014; Wang and Xiao, 2017). Table 2 summarizes the principal studies developed in the framework of SD forecasting by SVM methods. It is important to note that the studies compared different ML tools to find a reliable model for their defined objective.

Table 2. Applications of SVM in the field of sustainable development

Specific tools and applications in the machine learning model	Objective	Information regarding the training process	Model accuracy metrics	Study
Cobweb algorithm for hierarchical clustering to identify regions with similar indicator behavior in the framework of social, economic, environmental, and institutional dimensions, followed by an ordinal classifier: SVM or logistic regression (LR) SVM and LR classifiers and their reformulation to ordinal regression, were compared with their ensemble versions. The Gaussian function was applied as a kernel function	Classification and ranking countries according to its SD level.	Two methods developed for the training process: 1. A dataset with 75% of the patterns for training and 25% for testing, applied 30 times. 2. Three different data partitions were used with different years for training and testing.	MMAE= 0.28 Acc= 92.5 for the SVM trainable ensemble with ordinal coding	(Pérez-Ortíz et al., 2014)

Specific tools and applications in the machine learning model	Objective	Information regarding the training process	Model accuracy metrics	Study
Principal component analysis (PCA): for the dimension to reduce of environmental indicators, which was followed by SVR or ANN. 1. SVM-SVR, the Gaussian radial basis function was applied as a kernel function and, a sequential minimal optimization algorithm for data training. 2. ANN - radial basis neural network function (•)	Forecast model for eco-efficiency up to 1 year in advance.	A training dataset with 13 environmental indicators measured over 13 years and a validation dataset with 13 indicators measured for 1 year	RER= 1) 0.48%; 2) 2.9%	(Li et al., 2006)
Support vector spatial dynamic and mixed data sampling (SVSD-MIDAS) to consider spatial correlation, sampling data frequency, and the non-linear relationship between eco-efficiency and the factors. Additionally, the radial basis function kernels were adopted	Prediction of the regional eco-efficiency with spatial mixed-frequency panel data	A training and validation dataset with information from 1998 to 2013	Average MPE = <1% in different spatial settings Average MSE = <1% in different spatial settings	(Wang and Xiao, 2017)

The dots (•) in Table 2 identify the machine learning tool(s) that are not support vector machines. This type of ML was compared with the SVM employed in the study of interest.

- Artificial Neural Networks (ANN)

ANNs are non-linear operators formed by a set of neurons connected to each other and to an external environment through weight-determined connections (Sierra, 2006). They generalize non-linear relationship patterns between input and output variables with noise information (Kadiyala and Kumar, 2017b), and are used for industrial, energy, agriculture, environmental, transportation, economic and water conservation purposes (Chen et al., 2018). ANNs are composed of the following three layers: 1) an input layer that receives the predictor variables; 2) one or more hidden layers, which usually share the same information, and are interconnected in different ways; and 3) an output layer; for regression, the output layer may be a single layer while in a classification case the output will consist of a node of possible output classes (Holloway and Mengersen 2018). It is important to mention that each node or neuron in the hidden layers represents an activation function that acts on a weighted input of the previous layers' outputs (Holloway and Mengersen 2018). Neural networks can be classified based on their number of layers (monolayer or multilayer networks) and according to the direction that the information flows (recurrent networks, feedforward networks). The most widely known ANN is the multilayer perceptron (MLP). The back-propagation (BP) algorithm performs better the ANN given its capacity to identify the correct weight of the nodes in the ANN.

In SD modeling, artificial neural networks simulate non-linear relationships among indexes and prevent bias found in traditional weight design methods (Zhang et al., 2009). They have been used in evaluating SD by applying the BP algorithm to establish the degree of relationship between indexes (Zhang et al., 2009; Zhang and Huan, 2006), and to determine the behavior of the environmental dimension and activity levels in an emissions inventory through the use of sustainability indicators (Antanasijević et al., 2013). Table 3 outlines the aforementioned studies and indicates the reliability of the machine learning structures employed.

Deep neural networks (DNNs) are ANN, which cover complex architectures and have more than one layer of hidden units, which improves the computer's learning capacity. They are used in the energy field through multivariate analyses of environmental, social and economic factors with big data (Ifaei et al., 2017). They are also employed in the field of greenhouse gas mitigation, as a basis for developing SD plans (Ifaei et al., 2017;

Madu et al., 2017), and as a key tool for air quality forecasts that considers particulate matter (Karimian et al., 2019; Y. Zhou et al., 2019).

A little number of publications that employed ANN to predict sustainable development, or a combination of its dimensions, was found in the studies analyzed. Table 3 summarizes those studies and the application of ANN over the last 20 years, indicating their use and the resulting metrics used to evaluate the performance of the developed ML models.

Table 3. Applications of ANN in the field of sustainable development

Specific tools and applications in the machine learning model	Objective	Information regarding the training process	Model accuracy metrics	Study
BP neural network: definition of the degree of relationship among indexes in the same layer.	Evaluation of sustainable development levels for society, economy, zoology, resource, science, and technological factors	10 years of statistical data used in the evaluation (1993-2003)		(Zhang and Huan, 2006)
BP neural network to estimate the index weight of sustainable development and forecasting SD factors.	Forecasting sustainable development factors	A training dataset with 33 factors from 8 years, and another with 1 year of data for testing.	Average error 0.037	(Zhang et al., 2009)
Six annual hourly consumption variables predicted by recurrent and deep neural networks as machine learning tools. The prediction was part of the information of the Techno-Econo-Socio-Environmental Multivariate Analysis (TESEMA) model which additionally encompassed: 1. PCA: Identifying linear modeling between variables. 2. k-Nearest Neighbors (k-NN): Clustering zones according to modeling results. 3. Multivariate data analysis (TESEMA): Reducing data variability. 4. Partial least squares (PLS): investigate any correlation between major variables of TESEMA and population density	DNN: prediction of annual hourly electrical demand load using the maximum domestic power consumption at peak intervals, industrial power consumption, stored power at power plants, and total energy trade. TESEMA model: Analyzing the correlation between variables and population density in the study zone.	Two-thirds of available data as a training set and the rest as the testing set.	RSME = 73.15% for DNN	(Ifaei et al., 2017)

- Decision Trees

Decision trees (DT) are an inductive inference method of machine learning that develops classification rules or regression tasks. The former consists of a supervised learning method that uses class-labeled training examples; the leaves are the labeled classes, and the branches are the features that conduct the classification tasks. This kind of tree creates classification rules to be applied for new observations to be classified. Decision tree regression enables the forecasting of continuous variables based on input predictors. For example, decision trees can to be applied for pattern recognition with regard to specific characteristics of an area, which was the case in a study developed by Zeng et al. (2019), or to forecast pollution levels according to the behavior of pollution sources (Zalakeviciute et al. 2020). M5P, C4.5 and random forests are decision tree algorithms employed in studies reviewed. The C4.5 algorithm can be used for classification tasks and builds decision trees using the information gain concept (Tzima 2011), while the M5P algorithm combines trees with linear regression models

at the leaves (Shaban 2016). The C5.0 algorithm considers the information entropy to select the best classification method. To improve the robustness of the model, C5.0 introduces boosting technology that consists of repeating sampling simulation of existing weighted samples (Zeng et al. 2019).

Random forests (RF) correspond to a collection of DT that are applied to classification and regression problems (Brink et al., 2016). They make statistical predictions by averaging a set of non-correlated regression or classification trees, and are capable of computing non-linear relationships and interaction effects (Zhan et al., 2018). For example, they can be used to create potential land use transition maps by applying the RF classification algorithm to satellite images, combining dynamic, biophysical, socioeconomic and legislative factors (Gounaridis et al., 2018). Table 4 outlines the aforementioned studies and the application of DT over the last 20 years.

Table 4. Applications of decision trees in the field of sustainable development

Specific tools and applications in the machine learning model	Objective	Information regarding the training process	Model accuracy metrics	Study
Two decision tree models that applied the C5.0 growth and pruning algorithm for the 1) development stage target and 2) region target.	Identifying patterns and characteristics of sustainability of coal-mining cities considering data from 55 prefectures and 34 indicators	The development stage of the 55 coal-mining cities was the classification attribute	Acc= 92.73% for the stage model, Acc=94.55% for region model.	(Zeng et al. 2019)
1. Remote sensors and Landsat images: determining changes in urban dynamics. 2. Random forests: 3 predictor variables randomly sampled at each decision tree split and 500 trees to be built 3.Cellular automata: Forecasting land use changes through 2045	Development of potential transition surfaces by combining 20 dynamic, biophysical, socio-economic and legislative factors.	A set of randomly distributed points was plotted against the Landsat images and high-resolution images available via Google Earth.	Acc= 90.9-93.1% with respect to the resulting yearly map compared against reference samples	(Gounaridis et al., 2018)

Combining ML with other tools, such as those presented in Tables 2 – 4, has improved pattern recognition efficiency and data clustering to more accurately forecast the behavior of an analyzed data set.

3.1.3 Important variables that influence sustainable development

The relationship between the atmospheric component and SD has been demonstrated through the development of sustainable energy plans based on dimensional impact analyses and energy consumption (Ifaei et al., 2017), in addition to forecasting air quality by using certain sustainability indicators (Antanasijević et al., 2013). Studies predict and/or analyze atmospheric pollutant behavior based on information registered by monitoring stations and other sources, while advancing the integration of geospatial information, which includes the relationship between geospatial data and mixed frequency data analysis through SVM (Wang and Xiao, 2017). The same occurs with the application of the BP algorithm in ANNs, establishing the degree of relationship of indexes in the same layer (Zhang et al., 2009). The results of the relationship between different variables and information characteristics are emphasized in order to understand future land use behavior in a territory (Gounaridis et al., 2018).

The scope of most studies is both from a holistic vision of SD and the specific analysis of a variable or indicator in the framework of SD. There is a lack of studies that identify the most influencing variables or indicators of SD progress. Urban territories are influenced by different variables and identifying the progress of each and their synergetic behavior in the scope of SD is necessary for decision-makers. Identifying variables' level of

importance should consider the data that feeds the models. In this vein, tools like analytic hierarchy process and clustering could guide the knowledge of influencing variables in sustainability categories. Hybrids or ensembles of tools offer an improvement of the models' exactitude or reduced computational costs, as demonstrated in the studies developed by Peng et al. (2017) and Zhan et al. (2018). Therefore, the ensembles, which include hierarchical and clustering tools, could support the identification of important variables regarding the sustainable development behavior of territories, as well as its forecast.

Even though there are a limited number of studies associated with the forecast of SD and its influential variables, research for air quality forecasting through ML tools provides guidelines for causality analyses of urban sustainable development progress. Tools like proportion-based causality tests and sequential forward feature selection applied by (Wang, 2019) facilitate the evaluation and forecasting of urban sustainable development.

3.2 APPLICATION FOR AIR QUALITY

ANNs, SVMs and DTs have been utilized to forecast air quality and their influencing variables. Studies primarily focus on Asia and Europe. Countries such as India, China and Saudi Arabia, which according to the WHO have higher registered levels of pollution, stand out due to their application of ML tools to forecast future pollution events.

ANNs have been employed for air quality and atmospheric emission applications, with MLP and BP predominantly used as training algorithm. To identify the best forecasting tool, ANNs have been compared with conventional statistical models, DTs, and SVMs. Multilayer perceptron neural networks have been extensively analyzed and have been combined with the Manhattan propagation algorithm (Souza et al., 2015), with self-organizing map (SOM) and hierarchical clustering (Tamas et al., 2016), linear regression statistical techniques, or in comparison with an extreme learning machine (ELM) (Peng et al., 2017). Hybrid models provide synergetic results, in particular to establish early warnings, and have improved behavior in forecasting specific atmospheric pollutants. The application of ELM has surpassed the limitations of MLP with respect to the non-linearity of data and operational costs of traditional ANN methods (Peng et al., 2017).

To identify pollution risk, clustering algorithms such as the K-means algorithm have been used by applying the AQ algorithm to defined groups (Cervone et al., 2008). Furthermore, MLP was applied to forecast PM_{10} and PM2.5, while the K-mean and PCA algorithms facilitated the identification of input variables for the model (Franceschi et al., 2018).

For the ozone (O_3) and PM_{10} forecasts, lazy learning networks (LL) perform better than pruned neural networks (PNN) and feed-forward neural networks (FFNN), while PNNs effectively detect the increase of alarm and attention thresholds for the pollutants analyzed (Corani, 2005). To forecast the concentration of nanoparticles, FFNNs are combined with BP, in which the inclusion of all types of variables enabled the ANN to precisely map the non-linear relationship between the measured and expected nanoparticles (Al–Dabbous et al., 2017). However, to predict $PM_{2.5}$, it was concluded that increasing the number of data points does not necessarily result in better estimates, as it is more a correlation between the main factor and those related to it (Ni et al., 2017).

Additionally, the effectiveness of different ML algorithms was compared, as well as integration with the Gaussian dispersion model for an emission source, establishing excellent behavior in forecasting the dispersion of atmospheric pollutants, which is an applicable method for predicting and identifying source parameters (Ma and Zhang, 2016). Table 5 lists the results of the studies analyzed and how ANNs have been applied and combined with algorithms, see the artificial neural network model in Table 5 for model training and selecting input variables, while also presenting that a single study compared different tools.

Table 5. Application of ANNs to air quality and atmospheric pollutants

Study	Artificial neural network model	Objective	Information regarding the training process	Model accuracy metrics
(Souza et al., 2015)	MLP configured with hyperbolic tangent activation function + Manhattan propagation algorithm	Forecast daily concentrations of PM_{10}.	80% of data used for training (daily samples collected from (07/2009-06/2013) and 20% for testing the data set (daily samples collected from 07.2013-06.2014).	Improvements of MSE compare with individual MPL and ensembles = 8.85% (4 neurons in the hidden layer).
(Tamas et al., 2016)	1. MLP + Levenberg-Marquardt training algorithm + hybridized with hierarchical clustering. 2. MLP hybridized with SOM and k-mean clustering.	Hourly concentrations of O_3, NO_2 and PM_{10}, 24 hours in advance.	Clustering methods were used to subdivide the data set, with MLP trained on each subset: 60% for training data (3 years of data), 20% for validation (1 year of data), and 20% for testing (1 year of data).	RMSE =18.65 (O_3); 12.1 (NO_2); 7.4 (PM_{10}) μgm^3 MAE =14.69 (O_3); 8.58 (NO_2); 5.77 (PM_{10}) μgm^3 IA = 0.87 (O_3); 0.8 (NO_2); 0.74 (PM_{10})
(Peng et al. 2016)	ELM based on MLP+ hill-climbing algorithm to determine the optimal number of hidden nodes. ELM conducted the training task, followed by an online sequential ELM (OSELM) which updated the model. The input data of the model consisted of meteorological variables, O_3, $PM_{2.5}$, NO_2, and physical variables. An online sequential multiple linear regression and the MLP configured with hyperbolic tangent activation function were compared together with the OSELM.	Hourly concentrations of O_3, NO_2 and $PM_{2.5}$, 48 hours in advance.	2 years of information for training and validation (2009/07-2011/07), three for testing (2011/08-2014/07) as well as model updating.	The models were ranked using the forecast scores averaged over all forecast lead times, for each pollutant. For MAE and correlation scores the OSELM outperformed MLP and MLR for O_3, NO_2 and $PM_{2.5}$

Study	Artificial neural network model	Objective	Information regarding the training process	Model accuracy metrics
(Franceschi et al., 2018)	A combination of PCA to identify the most influencing predictors, K-means for data grouping, and MLP neural network + BP as the training algorithm.	Hourly and daily prediction of PM_{10} and $PM_{2.5}$.	Ratios for training and validating the data set: 80% for training and 20% for validation.	RSME= 15.62 (PM_{10}); 5.79 ($PM_{2.5}$) µg/m^3 MAE=,13.39 (PM_{10}); 4.72 ($PM_{2.5}$) µg/m^3
(Paas et al., 2017)	MLP configured with a hyperbolic tangent transfer function + BP training algorithm.	Prediction of ($PM_{0.25-10}$) mass concentrations and particle number concentrations.	Data split through stratified random sampling with self-organizing map. 70% of the subset was used for training, 20% for validation, and 10% for testing.	RMSE= 7.78 µg/m^3
(Ni et al., 2017)	Back propagation neural network.	A correlation analysis of $PM_{2.5}$, meteorological data, pollutant concentration data and social media.	Ratios for modeling and testing the data set: 70% for training and 30% for testing.	RMSE=24.06 µg/m^3
(Chen et al., 2018)	Measurement of partial mutual information to select significant variables + ensemble ANN-based output estimation + KNN regression output estimation error.	Forecasting an air quality index one day in advance considering the following predictors: PM_{10}, $PM_{2.5}$, and SO_2	2 years of information for training, one for validation.	The model cannot simulate well for large AQI values, but has good precision for medium and small AQI values
(Antanasijević et al., 2013)	General regression neural network + genetic algorithm for training.	Forecasting PM_{10} up to two years in the future with the following predictors: GDP, gross inland energy consumption, wood incineration factor, motorization rate, paper production, and processing of certain minerals.	5 years of data from 26 countries for training and validation, 2 years for testing.	MAE = 10%
(Corani, 2005)	1. FFNN configured with a hyperbolic tangent transfer function and the Levenberg-Marquardt training algorithm. 2. LL 3. PNN	Prediction of O_3 and PM_{10} at 9:00 am for the current day and detection of exceedance.	Cross-validation approach	True/predicted correlation=0.85 (O_3); 0.9 (PM_{10}) Success index=0.6 (O_3); 0.75 (PM_{10}) MAE=15.87 (O_3) and 8.25 for LL which outperformed the FFNN and PNN in the prediction.

Study	Artificial neural network model	Objective	Information regarding the training process	Model accuracy metrics
(Al–Dabbous et al., 2017)	FFNN configured with a hyperbolic sigmoid transfer function + BP training algorithm based on Levenberg Marquardt optimization.	Prediction of nanoparticles regarding different input variables.	Ratios for modeling and testing the data set: 80% for training and 20% for testing.	R^2 value = 0.79 IA = 0.94
(Ma and Zhang, 2016)	1. RBF + Gaussian dispersion model	Emission source parameters identification and pollutant dispersion forecasting.	Ratios for modeling and testing the data set: 47% for training and 53% for testing.	MSE=482.78 R=0.89
	2. BP + Levenberg-Marquardt training algorithm			MSE=371.82 R=0.91
	3. SVR+ Gaussian dispersion model			MSE=300.37 R=0.92
(Zhou et al., 2019)	1. Shallow multi-output long short-term neural network memory (SM-LSTM)	Regional multi-step-ahead air quality ($PM_{2.5}$, PM_{10}, NO_x) forecast horizon: t+1 up to t+4	Mini-batch gradient decent algorithm, dropout neuron algorithm, and L2 regularization algorithm for the training process.	MSE= 0.87
	2. Deep multi-output LSTM (DM-LSTM) neural network			MSE= 0.72
(Karimian et al., 2019)	1. Long short-term neural network (LSTM)	Prediction of $PM_{2.5}$ up to 48 hours.	Ratios for training and validating the data set: 60% for training, 20% for validation, and 20% for testing.	RMSE=8.91µg/m^3 MAE= 6.21 µg/m^3 R^2= 0.8
	2. DNN: Deep feed forward neural network (DFNN)			RMSE=19.45 µg/m^3 MAE= 14.52 µg/m^3 R^2= 0.49
	3. Multiple additive regression trees (•)			RMSE=12.83µg/m^3 MAE= 8.99 µg/m^3 R^2= 0.56

The dots (•) in Table 5 identify the machine learning tool(s) that are not artificial neural networks. This type of ML was compared with the ANN employed in the study of interest.

MLP was compared with SVR, in which the generalization capacity acquired in a relatively small amount of learning data and a large number of entry nodes demonstrated better behavior for SVR (García et al., 2013; Shaban et al., 2016; Suárez et al., 2011). This capacity was also demonstrated in forecasting air quality indexes (Liu et al., 2017; Weizhen et al., 2014) (see Table 6). Furthermore, SVR was utilized for spatio-temporal modeling of PM2.5, taking into account missing information (Song et al., 2014), and to forecast atmospheric pollutants by developing an online method (Wang et al., 2008). SVR and the Gaussian function kernel have been used to capture the non-linearity and interaction between predictors (de Hoogh et al., 2018; Liu et al., 2017; Song et al., 2014; Wang et al., 2008; Weizhen et al., 2014). In the prediction of hourly and daily levels of carbon monoxide (CO), the hybrid model of SMV and partial least squares (PLS), which was used to reduce input data, performed better with less modeling time and better performance metrics than the just SVM (Yeganeh et al., 2012).

Table 6. Application of SVMs to forecast air quality and atmospheric pollutants

Study	SVM model	Objective	Information regarding the training process	Model accuracy metrics
(García et al., 2013; Suárez et al., 2011)	1. MLP neural network configured with a sigmoid activation function (•) 2. SVR + sequential minimal optimization algorithm, and kernel function variants	Obtain a relationship between concentrations of CO, SO_2, NO, NO_2, PM_{10}, O_3	10-fold cross-validation	SVM outperformed MLP with correlation coefficients that ranged from 0.62 to 0.9 for the pollutants.
(Liu et al. 2017)	SVR	AQI forecasting up to 24 in advance, with information ($PM_{2.5}$, PM_{10}, SO_2, CO, NO_2 and O_3 levels, AQI, temperature, weather, wind force and direction) on cities with similar urban air pollution.	4 folds, each with 25% of the data. 3 folds were selected for training, the reaming fold was used for model testing.	RMSE = 6.54 MAPE = 0.0534
(Weizhen et al. 2014)	SVR + successive over relaxation algorithm + Gaussian kernel function	Prediction of PM_{10} and $PM_{2.5}$ for the following 24h and hourly forecasting based on daily average aerosol optical depths and meteorological parameters.	k-fold cross validation. 83% of the data used as a training dataset, the remaining as a testing dataset.	R^2=0.87 Average error=12.66 ug/m^3
(Song et al., 2014)	Spatial data aided incremental SVR	Daily average spatio-temporal prediction of $PM_{2.5}$ based on hourly PM_{10} levels.	The split for training and testing the ML model encompasses data from August 2006 to 2009 for training and data from 2011 to 2012 as a testing dataset.	RMSE= 1.0775 MAE = 0.81 MBE = 0.18 IA= (0.51-0.68)
(Wang et al., 2008)	SVR model + Gaussian kernel function	Prediction of respirable particulate matter, NO_X and SO_2 concentrations, up to 24 hours and 1 week in advance by using data found online.	59% for training and 41% for testing the dataset.	RMSE=25.89ug/m^3 MAE=19.29 ug/m^3
(Yeganeh et al., 2012)	Partial least square (PLS) for data selection and to reduce the amount of input data + SVR + radial-basis function as a kernel function.	Hourly and daily CO concentration forecast by using data on PM_{10}, total hydrocarbons (THC), NOx, CH_4, SO_2, O_3 and meteorological parameters.	75% for training and 25% for testing the model.	RMSE = 0.711 R^2 =0.654

The dots (•) in Table 6 identify the machine learning tool(s) that are not support vector machine. This type of ML was compared with the SVM employed in the study of interest.

On the other hand, the M5P decision tree algorithm outperformed both SVM and ANNs, given the efficiency of its tree structure and generalization capacity (Shaban et al., 2016) (see Table 7). The M5P algorithm

performed well in forecasting PM_{10}, when heuristic rules were applied (Oprea et al., 2016). A comparable situation was found when applying the RF algorithm, which performed better than the SVM algorithms (Pandey et al., 2013). RFs outperform other classifiers because of their capacity to assimilate forecasts from a variety of simple tree classifiers based on more predictive variables, resulting in low levels of bias and variance (Pandey et al., 2013). By applying different predictors, it was found that different variables have different types of impacts on the levels of ultrafine particulate matter in the atmosphere. The same case was presented in a comparison of five types of algorithms: LR, MLP, DT (C4.5 algorithm), SVM, and a variant of "ZCS-DM" decision trees, which belong to a class of ML tools, termed learning classifier systems, which use conditional rules. SVM and ZCS-DM performed well, the latter algorithm identified extreme cases and forecasted pollution episodes (Tzima et al., 2011).

SVM has been used to compare ML algorithm performance. SVM was compared with single decision trees (SDT), decision tree forests (DTF), and decision tree boosting (DTB) for forecasting air quality indexes (Singh et al., 2013). It was concluded that as the result of incorporating aggregation and optimization algorithms, DTF and DTB outperformed SVM both in classification and regression (Singh et al., 2013). RFs outperform chemical transport models, which require input variable information from emission inventories that may not be accurate (Zhan et al., 2018).

DTB was used to forecast PM_{10} concentrations, revealing that when comparing the multiple linear regression model (MLRM), the quantile regression model (QRM), and the generalized additive model (GAM), the capacity of the QRM to capture contributions from covariates in different quantiles produces better forecasting, when compared to procedures in which a single central trend is considered for a set of independent variables (Sayegh et al., 2014).

To improve the exactitude of the models and reduce computational costs in data treatment, analysis and forecasting, hybrids or ensembles of tools have been made (Peng et al., 2017; Zhan et al., 2018). The choice was made to combine DTs in their different degrees of complexity, which have been compared to the performance of conventional statistical tools, SVMs and ANNs. Furthermore, DTs have been combined with linear regression tools to improve their accuracy. Table 7 presents the studies analyzed, including the decision tree model developed in each study, the objective of the ML model, information regarding the training and testing datasets, and each model's performance. Some studies evaluated the performance of the ML model in comparison with other learnings architectures.

Table 7 Application of decision trees to air quality and/or atmospheric pollutants

Study	Decision tree model	Objective	Information regarding the training process	Model accuracy metrics
(Zhan et al., 2018)	RF	Prediction of the spatio-temporal distribution of daily 8h maximums of O_3 concentrations.	10-fold cross-validation (9 groups for training and 1 for testing).	RMSE=26 μg/m^3 R^2 = 0.69
(Sayegh et al., 2014)	Boosted regression trees	Prediction of hourly PM_{10} concentration levels.	10-fold cross validation; 11 months of data used as a training data set; 1 month as a testing dataset.	MBE=-41.1 μg/m^3 RMSE=125.6 MAE= 80.4 R=0.54 IA=0.66
(Singh et al., 2013)	1. Decision tree forest	• Seasonal air quality discriminatio n • Air quality index prediction	Kennard-Stone approach for a uniform scatter of training data around the training domain (70% for training and 30% for testing), k-folds-cross validation.	Acc=96.68% RMSE=6.58 MAE=5.24 R=0.90
	2. Decision tree boosting			Acc=96.45% RMSE=6.59 MAE=5.26

Study	Decision tree model	Objective	Information regarding the training process	Model accuracy metrics
				R=0.90
(Tzima et al., 2011)	3. Tree induction algorithm C4.5	Air quality episode forecasting	10-fold cross validation	Overall average rank regarding the kappa coefficient for pollutants=4.25 (C4.5); 2.5 (ZCS-DM); 1.75 (SVM); 3.4(MLP)
	4. Rule induction ZCS-DM algorithm			
	5. SVM (•)			
	6. MLP (•)			
(Shaban et al., 2016)	1. M5P decision tree	Prediction of NO_2, SO_2, and O_3 concentrations up to 24 hours in advance.	Time windowing (forecasting horizon) = 2 windows are used for training and testing.	RMSE=5.8
	2. SVM (•)			RMSE=6.4
	3. ANN (•)			RMSE=16.4
(Oprea et al., 2016)	1. M5P decision tree	Prediction of PM_{10} concentrations levels up to 3 days in advance.		RMSE=8.38; MAE=6.52; R=0.87
	2. Reduced error pruning tree algorithm			RMSE=12.74; MAE=9.34; R=0.65
(Wang, 2019)	1. DT	Air quality classification	k-cross validation	Acc= 75.1%
	2. Ensemble (boosted and bagged trees)			Acc= 90.2% and 89.1%
	3. k-NN			Acc= 80.2%
	4. SVM (linear and gaussian) (•)			Acc= 82.3% and 85.6%

The dots (•) in Table 7 identify the machine learning tool(s) that are not decision trees. This type of ML was compared with the DT employed in the study of interest.

The machine learning tools presented above enabled the forecasting of air quality or specific indicators (see Tables 5 – 7). Their performance and accuracy primarily depend on the data set and feature selection. Additionally, the machine learning tools listed in Tables 2 – 4 showed the first advances related to forecasting sustainable development or its dimensions. Nevertheless, hybrid or ensembles like those used by Karimian et al. (2019) and Wang (2019) improve specific ML tools, and could be applied in forecasting sustainable development, possibly even to identify influencing variables.

However, a specific analysis of the influence of air quality or atmospheric pollution on sustainable development was not found in any of the studies analyzed; nor a study that identifies the degree of importance of variables on sustainable development to achieve the targets defined in the SDGs. Specific aspects and results of the reviewed studies provided an understanding that there are two fundamental factors that need to be addressed in order to advance sustainable development forecasting; the information required for training and model validation, and the identification of important variables that influence sustainable development.

3.3 GAPS AND FUTURE DEVELOPMENTS

The application of ML must consider the characteristics of the data and its behavior. The atmosphere and reactions within it are non-linear complex processes, whose analysis is necessary to forecast SD and understand its impact. As such, SVMs and DTs are tools that can support this analysis.

The application of DNNs could be explored in conjunction with tools that have overcome data management difficulties. The majority of studies reviewed perform an analysis of redundancies to include variables with the most forecasting potential, avoiding the use of variables that introduce errors to the modeling results, which include over fitting.

One of the objectives in an air quality assessment is determining the environmental quality and reflecting human demands on the same (Chen et al., 2018). Some areas have geographic and atmospheric conditions that are not conducive to adequate air-circulation, coinciding with populations located in zones that are characterized by cities that are not optimally designed (Saeed et al., 2017). Given their effect on a population's health, these factors must be analyzed as they are part of the dynamic of SD.

As ultrafine particles are atmospheric pollutants that are not widely monitored, it is possible to learn their impact on the health of a population and on SD by applying ML and correlating them to other atmospheric pollutants. Machine learning tools, their ensembles, and using techniques such as analyzing satellite imaging, offer decision-makers the possibility of having a better understanding of their territories. In the framework of sustainable development and its targets, decision-makers everywhere have the responsibility of identifying the most effective actions to achieve the SDGs. The economic, environmental, and social dimensions are not static and require continuous support to maintain or improve their behavior. For example, any change to the environment and health dimensions of a population, both in an urban territory or country, could influence every sustainability dimension. Furthermore, the use of ensembles of machine learning tools that include hierarchical and clustering tools applied by Franceschi et al. (2018) and Tamas et al. (2016) with MLP neural networks or by Oprea et al. (2016) and Singh et al. (2013) who used different kinds of decision trees, or by Peng et al. (2017) who applied SVR to identify subsets of the most relevant attributes for the predictions task, may facilitate early decisions from leaders.

Several studies have successfully applied machine learning tools to forecast air quality, with accuracies ranging from 70-95%. In this sense, the methodology and ML tools described by the studies mentioned above are a reference for the analysis of sustainability and its forecasting. Furthermore, research studies developed by (Wang, 2019) outline the relevant methods and tools replicable to evaluate the influence of air quality on urban sustainable development.

4 CONCLUSIONS

Forecasting the level of SD and its variation in function of certain input parameters, enables decision-makers to establish the emphasis variable for increasing the sustainability of an analyzed territory. Through the analysis of the references consulted, the use of ML in defining sustainability indexes was evident, supporting the relevance of its application, as biases are prevented in determining the weights of parameters and indicators.
Regarding the research questions of this study, it is important to note that SVMs, ANNs, RFs, and DNNs were the machine learning tools used either to forecast sustainable development or their dimensions in different territories. The most widely used tools in the field of SD are SVMs and to a lesser degree, ANNs and DTs. However, it is important to note that no study was found which identified the influence of air quality on sustainable development. Nevertheless, different approaches and distinct machine learning tools have been applied, which enable further research to make progress in determining the influence of air quality on sustainable urban development, as well as in its forecasting. In this vein, combining ML with other tools, such as hybrid systems or ensembles that include hierarchical and clustering tools, is useful in identifying the influence of different variables on urban sustainable development.

With respect to ML tools and the non-linearity of information that characterizes the dimensions of sustainable development, SVMs and DTs have performed the best in computing this type of data. However, despite the existence of ML approaches with certain characteristics that makes them suitable for forecasting, boundary conditions establish an opportunity for their use. It is necessary to consider the tools' favorable aspects, applying those that adjust to specific conditions, including the availability of information with the required continuity and quality that enables the proper reflection of behavioral patterns of variables in a given territory, which is the key to forecasting.

Only the studies that applied ML in the field of air quality established the required time for forecasting or data treatment as an analytical value. This is a possible parameter of interest given the importance of making effective use of time in the development of these types of studies.

This study is useful in the field of environmental, social and economic dimension analysis, as it provides a map of different ML tools and the authors who have used them in specific case studies, either to address a specific challenge or in integrating the dimensions.

There is a need for studies that identify the behavior of territories concerning indicator variation and sustainable development variables. The environmental, social, and economic dimensions are interconnected, in which each all have an effect on the others. Furthermore, to understand and identify the most influencing actions on health, the environment, and economic effects, it is necessary to forecast behaviors and identify influencing variables, to develop scenarios and make the best decisions possible. The environmental dimension affects the social dimension in the framework of the population's health.

This study is unique and innovative, it recognizes the importance of air quality in sustainable development and seeks to identify the behavior of the ML tools applied in different studies to forecast both concepts. It also identifies the tools used to understand the influence of variables on sustainable development or the dimensions of sustainability, in addition to those used to establish the association of predictors. Even when documentation was found regarding machine learning tools, it is necessary to collect the experiences developed for decision making. Few studies have taken the initiative of forecasting sustainable development and analyze its most-influencing possible variables. This work identifies the degree of progress in this sense, as it supports decision-making that could be undertaken by governments as part of their goals and objectives and those tied to the 2030 Agenda for Sustainable Development.

5 REFERENCES

Al–Dabbous A, Kumar P, Khan A (2017) Prediction of airborne nanoparticles at roadside location using a feed–forward artificial neural network. Atmos Pollut Res 8:446–454. https://doi.org/10.1016/j.apr.2016.11.004

Antanasijević D, Pocajt V, Povrenović D, Ristić M, Perić-Grujić A (2013) PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci Total Environ 443: 511–519. https://doi.org/10.1016/j.scitotenv.2012.10.110

Brink K, Richards J, Fetherolf M (2016) Real-world machine learning. Richards & Fetherolf, Eds. Manning Publications.

Cervone G, Franzese P, Ezber Y, Boybeyi Z (2008) Risk assessment of atmospheric emissions using machine learning. Nat Hazard Earth Sys 8: 991–1000. https://doi.org/https://doi.org/10.5194/nhess-8-991-2008

Chen S, Kan G, Li J, Liang K, Hong Y (2018) Investigating China's Urban Air Quality Using Big Data, Information Theory, and Machine Learning. Pol J Environ Stud 27:565–578. https://doi.org/10.15244/pjoes/75159

Corani (2005) Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecol Model 185:513–529. https://doi.org/10.1016/j.ecolmodel.2005.01.008

Cruz C, Gómez A, Ramírez L, Villalva A, Monge O, Varela J, Quiroz J, Duarte H (2017) Calidad del aire respecto de metales (Pb, Cd, Ni, Cu, Cr) y relación con salud respiratoria: Caso Sonora, México. Rev Int Contam Ambie 33:23–34. https://doi.org/10.20937/RICA.2017.33.esp02.02

de Hoogh K, Héritier H, Stafoggia M, Künzli N, Kloog I (2018) Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland. Environ Pollut 233:1147–1154. https://doi.org/https://doi.org/10.1016/j.envpol.2017.10.025

Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. Atmos Pollut Res 9:912–922. https://doi.org/10.1016/j.apr.2018.02.006

García N, Combarro E, del Coz J, Montañes E (2013) A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study. Appl Math Comput 219:8923–8937. https://doi.org/10.1016/j.amc.2013.03.018

Gibert K, Sànchez-Màrre M, Sevilla B (2012) Tools for Environmental Data Mining and Intelligent Decision Support. In iEMSs. Leipzig, Germany. http://www.iemss.org/society/index.php/iemss-2012-proceedings. Accessed 26 November 2018

Gibert K, Sànchez-Marrè M, Izquierdo J (2016) A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. Ai Commun 29:627–663. https://doi.org/10.3233/AIC-160710

Gounaridis D, Chorianopoulos I, Koukoulas S (2018) Exploring prospective urban growth trends under different economic outlooks and land-use planning scenarios: The case of Athens. Appl Geogr 90:134–144. https://doi.org/10.1016/j.apgeog.2017.12.001

Holloway J, Mengersen K (2018) Statistical machine learning methods and remote sensing for sustainable development goals: A review. Remote Sens 10:1-21. https://doi.org/10.3390/rs10091365

Ifaei P, Karbassi A, Lee S, Yoo Ch (2017) A renewable energies-assisted sustainable development plan for Iran using techno-econo-socio-environmental multivariate analysis and big data. Energ Convers Manage 153:257–277. https://doi.org/10.1016/j.enconman.2017.10.014

Kadiyala A, Kumar A (2017a) Applications of R to Evaluate Environmental Data Science Problems. Environ Prog Sustain 36:1358–1364. https://doi.org/10.1002/ep.12676

Kadiyala A, Kumar A (2017b) Vector Time Series-Based Radial Basis Function Neural Network Modeling of Air Quality Inside a Public Transportation Bus Using Available Software. Environ Prog Sustain 36:4–10. https://doi.org/10.1002/ep.12523

Karimian H, Li Q, Wu Ch, Qi Y, Mo Y, Chen G, Zhang X, Sachdeva S (2019) Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations. Aerosol Air Qual Res 19:1400–1410. https://doi.org/10.4209/aaqr.2018.12.0450

Krzyzanowski M, Apte J, Bonjour S, Brauer M, Cohen A, Prüss-Ustun A (2014) Air Pollution in the Mega-cities. Current Environmental Health Reports 1:185–191. https://doi.org/10.1007/s40572-014-0019-7

Lässig Kersting and Morik (2016). Computat Sustainability. Springer. https://doi.org/10.1007/978-3-319-31858-5

Li Y, Wu Y.-X, Zeng Z.-X, Guo L (2006) Research on forecast model for sustainable development of Economy-Environment system based on PCA and SVM. In Proceedings of the 2006 International Conference on Machine Learning and Cybernetics (Vol. 2006, pp. 3590–3593). Dalian, China: IEEE. https://doi.org/10.1109/ICMLC.2006.258576

Liu B.-Ch, Binaykia A, Chang P.-Ch, Tiwari M, Tsao Ch.-Ch (2017) Urban air quality forecasting based on multi- dimensional collaborative Support Vector Regression (SVR): A case study of Beijing- Tianjin-Shijiazhuang. Plos One 12:1–17. https://doi.org/10.1371/journal.pone.0179763

Lubell M, Feiock R, Handy S (2009) City adoption of environmentally sustainable policies in California's Central Valley. J Am Plann Assoc 75:293–308. https://doi.org/10.1080/01944360902952295

Ma D, Zhang Z (2016) Contaminant dispersion prediction and source estimation with integrated Gaussian-machine learning network model for point source emission in atmosphere. J Hazard Mater 311:237–245. https://doi.org/10.1016/j.jhazmat.2016.03.022

Madu C, Kuei N, Lee P (2017) Urban Sustainability Management: A Deep Learning Perspective. Sustain Cities Soc 30:1–17. https://doi.org/10.1016/j.scs.2016.12.012

Mellos K (1988) Theory of Eco-development. In Perspectives on Ecology (pp. 59–74). London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-19598-5_4

Ni X.Y, Huang H, Du W.P (2017) Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. Atmos Environ 150:146–161. https://doi.org/10.1016/j.atmosenv.2016.11.054

Oprea M, Dragomir E, Popescu M, Mihalache S (2016) Particulate Matter Air Pollutants Forecasting using Inductive Learning Approach. Rev Chim 67:2075–2081. Retrieved from http://www.revistadechimie.ro

Paas B, Stienen J, Vorländer and Schneider Ch (2017) Modelling of Urban Near-Road Atmospheric PM Concentrations Using an Artificial Neural Network Approach with Acoustic Data Input. Environments 4:1-25. https://doi.org/10.3390/environments4020026

Pandey G, Zhang B, Jian L (2013) Predicting submicron air pollution indicators: a machine learning approach. Environ Sci- Proc Imp 15:996–1005. https://doi.org/10.1039/c3em30890a

Peng H, Lima A, Teakles A, Jin J, Cannon A, Hsieh W (2017) Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. Air Qual Atmos Health 10:195–211. https://doi.org/10.1007/s11869-016-0414-3

Pérez-Ortíz M, de La Paz-Marín M, Gutiérrez P.A, Hervás-Martínez C (2014) Classification of EU countries' progress towards sustainable development based on ordinal regression techniques. Knowl-Based Syst 66:178–189. https://doi.org/10.1016/j.knosys.2014.04.041

Phillis Y, Kouikoglou V, Verdugo C (2017) Urban sustainability assessment and ranking of cities. Comput Environ Urban 64:254–265. https://doi.org/10.1016/j.compenvurbsys.2017.03.002

Saeed S, Hussain L, Awan I, Idris A (2017) Comparative Analysis of different Statistical Methods for Prediction of PM2.5 and PM10 Concentrations in Advance for Several Hours. Int J Comput Sci Netw Secur 17:45–52.

Sayegh A, Munir S, Habeebullah T (2014) Comparing the Performance of Statistical Models for Predicting PM10 Concentrations. Aerosol Air Qual Res 14:653–665. https://doi.org/10.4209/aaqr.2013.07.0259

Shaban K, Kadri A, Rezk E (2016) Urban Air Pollution Monitoring System With Forecasting Models. IEEE SENSORS JOURNAL, 16:2598–2606. https://doi.org/10.1109/JSEN.2016.2514378

Sierra B (2006) Aprendizaje automático conceptos básicos y avanzados Aspectos prácticos utilizando el software Weka. Madrid, España: Madrid Pearson Prentice Hall.

Singh K, Gupta S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos Environ 80:426–437. https://doi.org/10.1016/j.atmosenv.2013.08.023

Song L, Pang S, Longley I, Olivares G, Sarrafzadeh A (2014) Spatio-temporal PM2.5 Prediction by Spatial Data Aided Incremental Support Vector Regression. In International Joint Conference on Neural Networks (pp. 623–630). Beijing: IEEE. https://doi.org/10.1109/IJCNN.2014.6889521

Souza R, Coelho G, da Silva A, Pozza S (2015) Using Ensembles of Artificial Neural Networks to Improve PM10 Forecasts. Chem Eng Trans 43:2161–2166. https://doi.org/10.3303/CET1543361

Suárez A, García P.J, Riesgo P, del Coz J.J, Iglesias-Rodríguez F.J (2011) Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). Math Comput Modell 54:453–1466. https://doi.org/10.1016/j.mcm.2011.04.017

Tamas W, Notton G, Paoli C, Nivet M, Voyant C (2016) Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. Aerosol Air Qual Res 16:405–416. https://doi.org/10.4209/aaqr.2015.03.0193

Toumi O, Le Gallo J, Ben Rejeb J (2017) Assessment of Latin American sustainability. Renew Sust Energ Rev 78:878–885. https://doi.org/10.1016/j.rser.2017.05.013

Tzima F, Mitkas P, Voukantsis D, Karatzas K (2011) Sparse episode identification in environmental datasets: The case of air quality assessment. Expert Syst Appl 38:5019–5027. https://doi.org/10.1016/j.eswa.2010.09.148

United Nations, Department of Economic and Social Affairs (2019). World urbanization prospects The 2018 Revision. New York. https://doi.org/10.18356/b9e995fe-en

Wang B (2019). Applying machine-learning methods based on causality analysis to determine air quality in China. Pol J Environ Stud 28:3877–3885. https://doi.org/10.15244/pjoes/99639

Wang W, Men C, Lu W (2008) Online prediction model based on support vector machine. Neurocomputing 71:550–558. https://doi.org/10.1016/j.neucom.2007.07.020

Wang X, Xiao Z (2017) Regional eco-efficiency prediction with Support Vector Spatial Dynamic MIDAS. J Clean Prod 161:165–177. https://doi.org/10.1016/j.jclepro.2017.05.077

WCED (1987). Report of the World Commission on Environment and Development: Our Common Future: Report of the World Commission on Environment and Development. Oslo. https://doi.org/10.1080/07488008808408783

Weizhen H, Zhengqiang L, Yuhuan Z, Hua X, Ying Z, Kaitao L, Donghui L, Peng W, Yan M (2014) Using support vector regression to predict PM10 and PM2.5. In IOP Conference Series: Earth and Environmental Science (Vol. 17). IOP. https://doi.org/10.1088/1755-1315/17/1/012268

WHO (2016). OMS | La OMS publica estimaciones nacionales sobre la exposición a la contaminación del aire y sus repercusiones para la salud. WHO. http://www.who.int/mediacentre/news/releases/2016/air-pollution-estimates/es/Accesed Accessed 26 November 2018

Yeganeh N, Shafie M. P., Rashidi Y, Kamalan H (2012) Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. Atmos Environ 55: 357–365. https://doi.org/10.1016/j.atmosenv.2012.02.092

Zalakeviciute R, Bastidas M, Buenaño A, Rybarczyk Y (2020) A Traffic-based method to predict and map urban air quality. Appl. Sci. 10. https://doi.org/10.3390/app10062035

Zeng L, Guo J, Wang B, Lv J, Wang Q (2019) Analyzing sustainability of Chinese coal cities using a decision tree modeling approach. Resour. Policy 64, 101501. https://doi.org/10.1016/j.resourpol.2019.101501

Zhan Y, Luo Y, Deng X, Grieneisen M, Zhang M, Di B (2018) Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ Pollut 233:464–473. https://doi.org/10.1016/j.envpol.2017.10.029

Zhang Y, Huan Q (2006) Research on the evaluation of sustainable development in Cangzhou city based on neural-network-AHP. In Proceedings of the Fifth International Conference on Machine Learning and Cybernetics (Vol. 2006, pp. 3144–3147). https://doi.org/10.1109/ICMLC.2006.258407

Zhang Y, Shang W, Wu Y (2009) Research on sustainable development based on neural network. In 2009 Chinese Control and Decision Conference (pp. 3273–3276). IEEE. https://doi.org/10.1109/CCDC.2009.5192476

Zhou Y, Chang F-J, Chang L-Ch, Kao I-F, Wang Y.S (2019) Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. J Clean Prod 209:134–145. https://doi.org/10.1016/j.jclepro.2018.10.243

Apéndice D.

Minería de texto y aprendizaje automático para identificar prioridades de desarrollo sostenible

Documento versión del autor presentado como resumen extendido respecto a la ponencia "Minería de texto y aprendizaje automático para identificar prioridades de desarrollo sostenible" realizada en el VII Congreso Colombiano y Conferencia Internacional de Calidad del Aire y Salud Pública CASAP. Barranquilla, 14-16 de agosto 2019. Publicado en: 2019 Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASP) por IEEE: DOI**:** 10.1109/CASAP.2019.8916682

Nidia Isabel Molina-Gómez[1], Camilo Andrés Rodríguez[1], P. Amparo López-Jiménez[2,] José Luis Díaz-Arévalo[3]

[1.]Facultad de Ingeniería Ambiental, Universidad Santo Tomás, Bogotá, Colombia
[2.]Departamento de Ingeniería del Agua y Medio Ambiental, Universitat Politècnica de València, Valencia, España
[3] Facultad de Ingeniería Civil, Universidad de la Sabana, Chía (Cundinamarca), Colombia

---

La aplicación de minería de texto a la información que se divulga a la comunidad en la web acerca de los objetivos del milenio y objetivos de desarrollo sostenible (ODS) permitió identificar los temas más frecuentes, así como su cohesión con los ODS. Los resultados del análisis fueron ejes de orientación y fundamentación para la inclusión de indicadores en el set de evaluación en el marco de la tesis doctoral.

RESUMEN

Alrededor del término desarrollo sostenible se han generado planes, programas y estrategias enfocados en la satisfacción de las necesidades de las generaciones del presente sin limitar la satisfacción de las necesidades de las generaciones futuras. Bajo este fin, los gobiernos a nivel mundial han realizado importantes esfuerzos, por un lado, para alcanzar los objetivos del milenio y ahora en el marco de los Objetivos de Desarrollo Sostenible (ODS). El gobierno colombiano asume el desafío de garantizar el logro de una vida sana para sus habitantes, la cual requiere de la interconexión e integralidad de los ODS. A nivel urbano se incluyen, entre otros desafíos, la mejora en los indicadores de salud, asociados a la contaminación del aire y que son alcanzables desde un marco de política y para su logro requiere de la participación de los grupos de interés encabezados por la comunidad. Este estudio analizó la información publicada en diversos medios de comunicación virtuales, del orden nacional, en el periodo 2009 a 2018; se aplicaron técnicas de minería de texto y herramientas de aprendizaje automático, mediante el uso del software R. Se identificaron temas centrales, sobre los cuales el país y su ciudad capital han prestado relevancia desde los medios de comunicación. Fue posible conocer el alcance en la divulgación de los retos, avances y oportunidades con la implementación de los Objetivos del Milenio, de los ODS y el papel de la calidad del aire.

Se encontró que el desarrollo sostenible no presenta mayor divulgación en los medios y la sostenibilidad se relaciona principalmente a la biodiversidad y las actividades financieras. Los esfuerzos se concentran en información específica y es escaza la divulgación de los retos, metas y avances en ODS. Aunque la calidad del aire es un tema de interés, no presentó un papel relevante como lo soportan los hechos de seguridad, pobreza, educación y calidad de vida. Un análisis de este tipo permite establecer las prioridades temáticas, de información y de desarrollo político tendencial, divulgados a la comunidad, como actor sobre quien recae la implementación de los lineamientos de política.

Palabras clave: ODS, ODM, minería de texto, comunicación, prioridades

1 INTRODUCCIÓN

La sostenibilidad en el desarrollo es un objetivo que se ha trazado desde 1987 con el establecimiento de su definición, pasando por la declaración realizada por las naciones unidas en el año 2000, acerca de la reducción de la pobreza y ocho objetivos a lograr para el año 2015 [1], los Objetivos de Desarrollo del Milenio (ODM): erradicar la pobreza extrema y el hambre; lograr la enseñanza primaria universal; promover la igualdad de género y la autonomía de la mujer; reducir la mortalidad infantil; mejorar la salud materna; combatir VIH/SIDA, paludismo y otras enfermedades; garantizar la sostenibilidad del medio ambiente y fomentar una asociación mundial para el desarrollo [1]. Posteriormente desde enero de 2016 se pusieron en marcha los 17 Objetivos de Desarrollo Sostenible (ODS), establecidos en virtud de los avances generados con los ODM y con la inclusión de nuevos objetivos, entre estos: cambio climático, la desigualdad económica, la innovación, el consumo sostenible y la paz y la justicia [2]. Estos objetivos son liderados a nivel gubernamental, estableciendo las acciones y presupuestos para lograr a 2030 las metas de desarrollo sostenible. Su implementación requiere de la participación del sector privado y del sector público; no obstante, los avances y resultados generados recaerán sobre la comunidad, quien del mismo modo podrá soportar su implementación.

Con el ánimo de identificar el grado de divulgación de los ODS y los anteriormente establecidos ODM, se analizó la documentación de artículos de prensa publicada en la web por medios de comunicación tradicionales de alto impacto; para ello se aplicaron técnicas de minería de texto a través del análisis semántico de textos no estructurados.

La minería de textos se ha aplicado en diversos estudios cuyo objetivo se concentra entre otros, en la identificación de énfasis o conjuntos temáticos predominantes en corpus textuales, así como en la identificación de patrones de una colección de textos. Entre la literatura revisada se encontró el desarrollo de estudios orientados a la identificación de vacíos de información y coocurrencias documentales, aplicados específicamente al análisis de planes de cambio climático [3], la identificación de tendencias y prácticas de sostenibilidad en procesos productivos [4], así como las tendencias y temas principales en publicaciones científicas [5]. El estudio más cercano al que se presenta en este trabajo aplicó técnicas de aprendizaje supervisado y procesamiento de lenguaje natural sobre documentos de noticias comparado con los indicadores reportados en informes de sostenibilidad [6]. Se encontró que, la minería de textos y el aprendizaje supervisado, permitieron identificar en las noticias los indicadores establecidos como importantes en los reportes de sostenibilidad, constituyéndose en una herramienta útil en la identificación de los problemas de sostenibilidad y de soporte a la identificación de indicadores específicos.

Los medios de comunicación son un canal directo con la comunidad, capaces de generar opiniones y de orientar la formación de conceptos basados en la información publicada. Son un grupo de interés que influencia a la comunidad sobre quien recaen los resultados en la implementación de los ODS y quien puede soportar las diferentes acciones establecidas en los planes de gobierno. En la revisión de literatura no se encontró la aplicación de herramientas de minería de texto a artículos de prensa para determinar el alcance y especificidad de la información publicada por los medios de comunicación en temáticas relacionadas con los ODM y ODS, mucho menos en temas específicos como lo es la calidad del aire, que se vincula al ODS: ciudades y comunidades sostenibles. Se evidenciaron dos estudios de análisis tendencial en la gestión de residuos [7,8] en que se analizaron textos en artículos de prensa y su comparación con el comportamiento de eventos del recurso hídrico.

A partir del análisis de la divulgación de los ODM y ODS en artículos de prensa publicados en la web, se buscó identificar, a través de este trabajo, los principales temas tratados en los últimos diez años (2009-2018) por los medios de comunicación, con el fin de identificar el grado de información que se entrega a la comunidad a propósito de los ODS y el papel de la calidad del aire como tópico fundamental en el objetivo ciudades y comunidades sostenibles.

2 METODOLOGIA

2.1 Recopilación de Información

Se desarrollaron cuatro pasos para cada año en el periodo de análisis y cada objetivo de desarrollo sostenible (ODS), consistentes en la búsqueda y recuperación de los artículos de prensa publicados en la web, su agrupación, limpieza preliminar de los textos recuperados y el almacenamiento de la información en documentos de formato txt.

Se generaron dos conjuntos de información. El primero, correspondiente a los artículos de prensa generados en Colombia y que incluyeran el término Colombia en su texto y, en el segundo conjunto, las noticias generadas en el país, que incluyeran a la ciudad capital: Bogotá. Para ambos conjuntos de datos se generaron dos subconjuntos: a) por año de emisión del artículo según ODS y b) por año e inclusión de las palabras clave: ODS, sostenibilidad y desarrollo sostenible. Para el primer subconjunto, se recuperaron artículos informativos en medios de comunicación de alto impacto en el país y a nivel nacional, mediante el uso de palabras clave, específicas para cada uno de los 17 ODS. En la tabla 1 se relacionan las palabras clave utilizadas para cada uno de los objetivos.

Tabla 1. Palabras clave según ODS para la búsqueda de información

ODS	Palabras Clave
1	pobreza, desastre, reducción de desastres
2	hambre, inseguridad alimentaria, malnutrición
3	mortalidad, vih, tuberculosis, malaria, hepatitis, cardiovascular, enfermedades respiratorias, diabetes, cáncer, suicidio, sustancias adictivas, accidentes de tráfico, planificación familiar, fecundidad, gastos sanitarios, contaminación ambiental, contaminación intradomiciliar, agua insalubre, saneamiento deficiente, higiene, intoxicación, consumo de tabaco
4	educación, tic, alfabetización
5	igualdad de género, violencia física, psicológica, sexual
6	agua potable, aguas residuales, tratamiento de aguas, uso eficiente del agua, estrés hídrico, saneamiento
7	energía, combustibles, electricidad, eficiencia energética
8	pib, empleo informal, desempleo, trabajo infantil, accidentes de trabajo, enfermedades laborales, derechos laborales, turismo
9	infraestructura resiliente, industrialización inclusiva, innovación, industrialización sostenible
10	desigualdad, discriminación
11	asentamientos informales, transporte público, crecimiento población, gestión urbana, preservación, protección y conservación del patrimonio cultural y natural, desastres, material particulado, partículas finas, contaminación del aire, zonas de esparcimiento, acoso, planes de desarrollo urbano y regional, reducción del riesgo de desastres, edificios sostenibles
12	consumo y producción sostenible, huella ecológica, pérdida de alimentos, desechos peligrosos, productos químicos, reciclaje, informes de sostenibilidad, turismo sostenible, combustibles fósiles
13	cambio climático, adaptación al cambio climático, gases de efecto invernadero, mitigación al cambio climático
14	eutrofización costera, acidez marina, conservación de mares, zonas protegidas, pesca sostenible, investigación, uso sostenible de los océanos
15	bosques, desertificación, pérdida de biodiversidad, degradación de tierras, gestión forestal sostenible
16	conflicto, muertes, violencia, armas, soborno, corrupción, secuestro
17	impuestos, asistencia oficial, inversión extranjera, ciencia y tecnología, internet

No se consideraron noticias publicadas en medios de comunicación internacional, aquellas que no integraran las palabras clave indicadas o que incluyeran información distractora no relacionada con los ODS.

El almacenamiento de la información se realizó en archivos de formato ".txt", de acuerdo con el año de emisión y por ODS, eliminando de los archivos la información publicitaria, las fechas de emisión de cada documento y el nombre de los redactores de los artículos de prensa. Para el ámbito nacional se recuperaron en total 7551

artículos noticiosos para un periodo de diez años comprendidos entre el 2009 a 2018, y 1976 para el ámbito distrital. Los artículos recopilados bajo las claves de búsqueda “ODS”, “sostenible” y “sostenibilidad” corresponden con un 6.3% para el caso nacional y 4% para la ciudad capital.

2.2 Procesamiento y Análisis de la Información

Mediante el uso del programa de código abierto R, se efectuó la preparación de texto de los documentos recopilados; se realizó limpieza a los textos eliminando términos que generaran ruido, palabras vacías, espacios vacíos, signos de puntuación y números; también se acotó el número de términos haciendo uso de sus raíces, mediante la lematización. Una vez preparado el texto se procede a la generación del corpus del documento y la creación de la nube de términos más frecuentes. Este procedimiento se aplicó para la información recopilada a) por ODS para todos los años, b) al conjunto de ODS por cada año y c) al total de documentos con los términos sostenibilidad, sostenible y ODS, en los conjuntos de información nacional y distrital.

Identificados los términos predominantes en cada conjunto y subconjunto de datos, se procede a establecer una matriz de frecuencias, se continúa con un análisis de asociación de términos frecuentes respecto a los demás términos del documento.

Adicionalmente, se hace un análisis de agrupación jerárquica de los términos con base en la distancia existente entre ellos. Se decidió utilizar la distancia “euclídea” y la distancia de “manhattan”; para ambos casos con el método de agrupación “promedio”, como método por defecto utilizado en el análisis de textos en el software R. La elección de la combinación de la distancia y método de agrupación se definió atendiendo al resultado del coeficiente cofenético que se presenta en un rango de 0 a 1, en donde la cercanía a 1 presenta un mejor comportamiento en la agrupación de los datos.

Finalmente, se generaron dendrogramas con la agrupación de los términos más frecuentes tomando la dispersión de términos en el rango del 90 a 97%, el factor de distancia de términos y el método de agrupación previamente definidos.

3 RESULTADOS

3.1 El enfoque de sostenibilidad

Del análisis de los artículos de prensa, recopilados bajo la clave de términos: sostenibilidad, ODS y desarrollo sostenible, se encontró que además de los términos clave de búsqueda, los más frecuentes para el ámbito nacional son: ambiente, social, empresa, agenda, ciudad, reto, paz. Términos en cuya agrupación, con base en una dispersión de 96%, se definen tres etiquetas: responsabilidad social, paz y agua (ver Fig. 1). Para el caso del componente atmosférico su aparición en el corpus del documento es reducida al encontrarse en la posición 46 de una matriz de 50 términos más frecuentes.

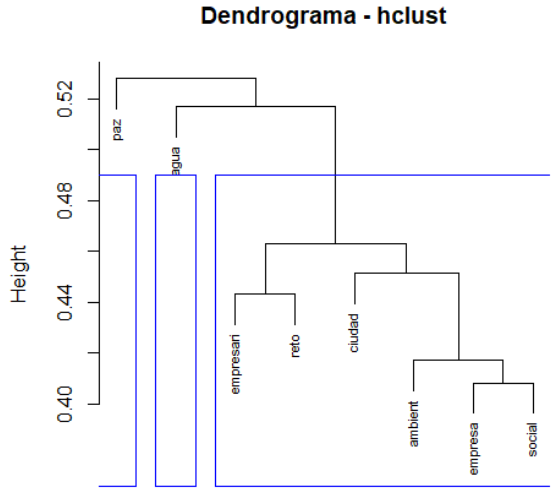


Fig. 1 Grupos de términos en los corpus asociados a sostenibilidad

Para el caso de Bogotá, a pesar de la relevancia en las claves de búsqueda, se encuentra dispersión en los términos del corpus documental, siendo las palabras más frecuentes: urbana y premio. Aunque los términos agua y aire son evidentes, se muestran en un mismo nivel que otros términos poco frecuentes. No fue posible aplicar un algoritmo de agrupación dada la alta dispersión de los términos en el conjunto de datos para el ámbito distrital.

3.2 ODS y sostenibilidad en el ámbito nacional y distrital

Al fusionar el conjunto de artículos de prensa clasificados según las palabras clave definidas por ODS con los artículos de prensa clasificados por los términos desarrollo sostenible, sostenibilidad y ODS, se encontró como términos frecuentes para el ámbito nacional: pobreza, agua y ambiente, seguido de los términos desarrollo sostenible, desempleo, mujer, desigualdad, hambre y violencia (ver Fig. 2). No es frecuente el uso de los términos ODS o milenio; tampoco se encontró alguna correlación del término "desarrollo sostenible" con los términos presentes en el corpus del documento analizado.

Fig. 2 Nube de términos relacionados con los ODS, y claves de búsqueda sostenibilidad, desarrollo sostenible y ODS en el ámbito nacional

Para el conjunto de todos los documentos en el ámbito nacional, con un coeficiente cofenético del 0.99, se encontraron tres etiquetas de términos: ambiente, desarrollo sostenible, agua-pobreza, se utilizó la distancia euclidiana, y dispersión del 95%. Para el caso de Bogotá se mantienen como términos frecuentes: agua, Transmilenio, plan, predial, reciclaje, arma, colegio y cerros (ver Fig. 3). Aunque el componente atmosférico no se observa como uno de los términos más frecuentes en el ámbito distrital, si presenta una asociación a términos como buses, contaminado, descontaminar, chimenea, integrado (relacionado al sistema de transporte de la ciudad) con un 57% de correlación para cada término.

Fig. 3 Nube de términos en el conjunto de ODS y claves de búsqueda sostenibilidad, desarrollo sostenible y ODS para Bogotá

4 ANALISIS DE RESULTADOS Y DISCUSION

Desde la definición de los ODM y ODS los estados han emprendido diversas acciones desde el ámbito gubernamental; se incluyen agendas de trabajo, productos técnicos con metas e indicadores que han sido divulgados desde dichas entidades a nivel documental y en sus portafolios web. Sin embargo, hace falta la integración de los medios de comunicación con información específica y relacional de los hechos que se enuncian como soporte a la comunidad, con la orientación a la generación de conceptos de manera más informada. Esto se evidencia en el enfoque central de los resultados encontrados para los últimos 10 años en los artículos de prensa en los ámbitos nacional y distrital.

Las clases generadas en el proceso de clúster son coherentes dentro de los ámbitos nacional y distrital analizados (ver Tabla 2). El ámbito de país presenta con mayor claridad las temáticas de algunos ODS, en comparación con el ámbito distrital.

Tabla 2 Resultados de agrupación de términos según corpus analizado: ODS y términos de sostenibilidad

Ámbito	Clases de términos	Coeficiente cofenético	Distancia	Dispersión de términos
Bogotá	1. arma 2. agua 3. plan cerros 4. predial 5. reciclaje 6. transporte	0,72	manhattan	97%
Colombia	1. ambiente 2. desarrollo sostenible 3. Agua- pobreza	0,99	euclídea	95%

De un análisis de la información publicada por entidades gubernamentales, acerca de indicadores relacionados a los ODM y ODS, se encontró que en un periodo de 6 años se redujo el indicador de pobreza en 12,6 puntos porcentuales [9]; estos años hacen parte del grupo de documentos analizados y los avances y actividades asociadas pudieron haber influenciado la información divulgada por los medios de comunicación.

Con relación al ODS 6, agua limpia y saneamiento, se reporta un incremento en 14,7 puntos porcentuales en tratamiento de aguas residuales en el país y una mejora en la calidad y continuidad en el servicio de acueducto y alcantarillado [9]. Aunque la gestión de las aguas residuales ha mejorado en el país, pasando del 27,8 % de aguas residuales urbanas tratadas en 2010 al 42,6 % en el año 2017 [10] el crecimiento poblacional continúa incrementándose y con ello el uso y aprovechamiento del recurso hídrico. Comunidades informadas acerca de estos indicadores, así como de su papel en la mejora de las condiciones son el soporte para el alcance de los ODS.

Adicionalmente la información que se suministra es específica de la situación territorial, no encontrando para Bogotá un enfoque claro en la información de los artículos de prensa acerca de la divulgación en materia de desarrollo sostenible, sostenibilidad, ODS y ODM. Dada la dispersión de los términos en el conjunto de datos no fue posible identificar grupos temáticos a partir de la agrupación no supervisada. Para el caso nacional se generaron tres grupos de términos; sin embargo, el término sostenibilidad, a la luz de la información que se divulga en los artículos de prensa en el país, está limitada a la responsabilidad empresarial y el recurso hídrico.

Para ningún conjunto de datos se generó un grupo temático relacionado con la calidad del aire. Sólo al analizar la información compilada en virtud de las palabras clave del ODS 11 ciudades y comunidades sostenibles, correspondiente al 5.7% del total de artículos recuperados y analizados, se encontró que los términos más frecuentes son: transporte, movilidad, ambiente y aire; además con una dispersión del 95% y un coeficiente cofenético del 0.86 estos mismos términos se presentan como etiquetas principales. No obstante, sólo se encuentra una asociación entre los términos aire y ambiente (46%). Además, pese a que la meta establecida para este ODS en materia de calidad del aire es *reducir el impacto ambiental negativo per cápita de las*

ciudades, incluso prestando especial atención a la calidad del aire … [11], no se encontró en el análisis de textos asociación a los términos gobernanza, políticas, planes, acciones o medidas; por su parte el término salud se observa de manera general al presentar correlación con términos como ambiente y ciudad.

Las prioridades de desarrollo sostenible identificadas se ven limitadas a las etiquetas generadas en el proceso de clasificación en cada conjunto y subconjunto de documentos analizado. Se trata de información que suministran los medios de comunicación acerca de hechos noticiosos y sobre los cuales la comunidad enfoca su atención. De un estudio realizado en España [12], se encontró que los jóvenes conceden a la información un alto valor cívico (8.2 en una escasa de 0 a 10); siendo importante pues posibilita y garantiza tanto el acceso al debate público como el desarrollo de una conciencia cívica [12].

Finalmente, al realizar una comparación con los resultados presentados por [6] y este estudio, se encontró que la información publicada por los medios de comunicación nacionales y distritales no es suficiente a la luz de identificar, rastrear y reportar indicadores de sostenibilidad. No obstante, coincide en que los medios de comunicación son un instrumento para identificar potenciales intereses de la comunidad y la ausencia de información en comparación con los retos y avances en materia de objetivos de desarrollo sostenible.

5 CONCLUSIONES

El componente atmosférico es un recurso vital, sobre el cual es reducida la información publicada en medios en los últimos 10 años. Llama la atención los términos más cercanos al recurso aire, en donde en su mayoría reportan a débiles condiciones de calidad de este, fuentes asociadas a la emisión de contaminantes atmosféricos, y orientaciones para su gestión que incluyen la reforestación.

En materia de ODS, Colombia revela información específica relacionada con agua-pobreza, ambiente, desarrollo sostenible. Para el periodo de análisis esta agrupación de términos, sugerida en la identificación de clases, mediante clúster, coincide con sólo algunos de los ODM y ODS. Por lo tanto, el papel de los medios de comunicación en la información completa requiere de revisión. Del mismo modo este trabajo se consolida en una oportunidad para aquellos medios de comunicación enfocados en la entrega de información ambiental, de responsabilidad empresarial y con intención de informar acerca del desarrollo sostenible. La información se fundamenta en la responsabilidad social y se desarrollan temas entorno a las necesidades suscitadas en el país.

Una etapa siguiente del presente estudio corresponde con la comparación de los resultados del presente estudio con las temáticas de ODS en redes sociales. Si bien es cierto las redes sociales y medios conversacionales son un medio preferido de información, son un vehículo a los medios tradicionales con los artículos de prensa. Adicionalmente se prevé una clasificación de documentos basado en las etiquetas generadas en este estudio.

6 REFERENCIAS

[1] UNDP, “Objetivos de Desarrollo del Milenio | UNDP.” [Online]. Available: https://www.undp.org/content/undp/es/home/sdgoverview/mdg_goals.html. [Accessed: 01-Jul-2019].

[2] PNUD, “Objetivos de Desarrollo Sostenible | PNUD.” [Online]. Available: https://www.undp.org/content/undp/es/home/sustainable-development-goals.html. [Accessed: 01-Jul-2019].

[3] M. W. Bickel, “A new approach to semantic sustainability assessment: text mining via network analysis revealing transition patterns in German municipal climate action plans,” 2017.

[4] W. Te Liew, A. Adhitya, and R. Srinivasan, “Sustainability trends in the process industries: A text mining-based analysis,” *Comput. Ind.*, vol. 65, no. 3, pp. 393–400, Apr. 2014.

[5] A. Schober, C. Kittel, R. J. Baumgartner, and M. Füllsack, “Identifying dominant topics appearing in the Journal of Cleaner Production,” *J. Clean. Prod.*, 2018.

[6] S. J. Rivera, B. S. Minsker, D. B. Work, and D. Roth, “A text mining framework for advancing sustainability indicators,” *Environ. Model. Softw.*, vol. 62, pp. 128–138, 2014.

[7] J. Wei, Y. Wei, A. Western, D. Skinner, and C. Lyle, “Evolution of newspaper coverage of water issues in Australia during 1843–2011,” *Ambio*, vol. 44, no. 4, pp. 319–331, May 2015.
[8] Y. Xiong, Y. Wei, Z. Zhang, and J. Wei, “Evolution of China’s water issues as framed in Chinese mainstream newspaper,” *Ambio*, vol. 45, no. 2, pp. 241–253, Mar. 2016.
[9] DNP, “Fin de la pobreza - La Agenda 2030 en Colombia - Objetivos de Desarrollo Sostenible,” 2019. [Online]. Available: https://www.ods.gov.co/es/objetivos/fin-de-la-pobreza. [Accessed: 07-Jul-2019].
[10] DNP, “CONPES 3948. Concepto favorable a la nación para contratar empréstitos externos con la banca multilateral o bilateral hasta por 40 millones de euros, o su equivalente en otras monedas, destinados al financiamiento parcial del programa para el saneamiento.” DNP, Bogotá D.C, pp. 1–38, 2018.
[11] PNUD, “Objetivo 11: Ciudades y comunidades sostenibles | El PNUD en América Latina y el Caribe.” [Online]. Available: http://www.latinamerica.undp.org/content/rblac/es/home/post-2015/sdg-overview/goal-11.html. [Accessed: 12-Mar-2018].
[12] A. Bergström and M. Jervelycke Belfrage, “News in Social Media,” *Digit. Journal.*, vol. 6, no. 5, pp. 583–598, May 2018.

Apéndice E.

Urban growth and heat islands: a case study in micro-territories for urban sustainability

Documento versión del autor sometido para posible publicación en la revista indexada en JCR "Urban Ecosistems" ISSN: 1083-8155/1573-1642 (Fecha en la que se sometió el documento manuscrito: 10.03.2021. Factor de impacto: 2.547, Q1 Scimago Journal Rank (SJR2019 0.87)

Nidia Isabel Molina-Gómez [a,b], Laura Marcela Varon-Bravo[a], Ronal Sierra-Parada [a], P. Amparo López-Jiménez[b]

[a] Department of Environmental Engineering, Universidad Santo Tomás, Carrera 9 51-11, 11321 Bogotá, Colombia
[b] Hydraulic and Environmental Engineering Department, Universitat Politècnica de València, Camino de Vera, 46022 Valencia, Spain

---

El uso de imágenes satelitales generó información relevante para el análisis del desempeño sostenible desde la dimensión ambiental. Además, se contrasta el comportamiento de la isla de calor urbano con las condiciones específicas del microterritorio.

Abstract

Rapid urbanization contributes to the development of phenomena such as climate variability in urban areas, especially in zones in tropical countries, which negatively impact ecosystems and humans, factors that influence urban sustainability. This growth generates considerable changes in green areas, which are replaced by surfaces. This is in addition to increased building construction that prevents the flow of wind streams contributing to the retention of pollutants and hot air masses, causing events such as urban heat islands (UHI). This study aimed to analyze from the micro-territorial level, the influence of urban growth on the UHI phenomenon over the last two decades (2000-2020). For this purpose, spectral indices calculated with satellite images were examined, in addition to socio-economic factors based on census data. The behavior of the following indices was analyzed: normalized difference vegetation (NDVI), normalized difference built-up index (NDBI), modified normalized difference water index (MNDWI), and the normalized difference impervious surface index (NDISI). Furthermore, population density and energy consumption were studied. Calculating the land surface temperature was performed by using thermal bands, which therefore enabled the creation of temperature profiles to verify its annual behavior. Lastly, a principal component analysis was carried out to understand and corroborate the behavior of the indices and UHI. This step made it possible to identify the contribution of the micro-territory to the principal components of UHI within the framework of urban sustainability. The spatio-temporal changes in UHI reveal a growing trend over time, especially in impermeable areas where several economic activities, vehicular traffic, and population density converge.

Keywords: Urban heat island, land surface temperature, spectral indices, remote sensors, impermeable surfaces, micro-territories.

1. Introduction

The world's population has increased exponentially, and this trend will continue, especially in areas such as Asia, Africa, and Latin America. It was estimated that in 2018, 55% of people lived in cities and this figure is projected to increase by 13% by 2050 (United Nations, 2019). Therefore, there will be increased population density, the expansion of settlements, land use difficulties, increased energy consumption, pollution, and modifications of cities' micro-climates, which include the phenomenon of urban heat islands (UHI) (Singh et al., 2017).

UHI refer to the temperature difference between urban and rural areas (Amanollahi et al., 2016; Estoque and Murayama, 2017; Oke, 1982), which is inevitable in cities due to urbanization processes that include surfaces originally covered by vegetation being replaced by infrastructures such as streets, houses, and buildings. Additionally, surfaces that were once permeable become impermeable, which favors an increase of land surface temperature (LST) (Carpio et al., 2020; Estoque and Murayama, 2017; Singh et al., 2017). The increase in LST entails greater energy demand, which exacerbates air pollution, cardiovascular and respiratory diseases, and impacts humans' quality of life (Bokaie et al., 2016; Liu et al., 2020; Senanayake et al., 2013; D. Zhou et al., 2019)..

UHI is an effect that goes against urban sustainability, which can be evaluated from factors that describe the environmental, social, and economic behavior of territories (Shen et al., 2013). The environmental dimension includes issues related to air quality, biodiversity, water, and soil resources. The social dimension encompasses population growth, health-related effects on the population, in addition to the social conditions of access to services related to urban expansion and densification. Lastly, the economic dimension entails infrastructure as support for territorial development and consumption patterns (United Nations, 2007). These are major interrelated themes which, in the context of this study, are part of the sustainable development goal known as sustainable cities and communities (UNDP, 2020).

Studying UHI requires a technical analysis from the perspective of temporal and spatial changes, in addition to knowledge of LST distribution in areas that have gone through urbanization processes. Moreover, areas with unusual temperatures must be identified. Given its capacity to map thermal distribution, satellite image processing is used, as it enables the analysis of distributed LST at the spatio-temporal level (Senanayake et al., 2013).

This phenomenon has been studied primarily in countries such as the United States, Germany, Greece, France, as well as in Asian countries such as China, India, and Japan, which have provided significant research on heat factors (Ulpiani, 2021; D. Zhou et al., 2019). Nevertheless, there is a lack of research on UHI in Latin America (Dobbs et al., 2018; Litardo et al., 2020; Peres et al., 2018; Portela et al., 2020; Wu et al., 2019), which is necessary due to continuous urbanization processes and increasingly intensifying climate sensitivity. Although urbanization processes generate effects in any territory, it is important to note that the urban growth rate in developing countries such as those in Latin America is 2.29% per year, compared to 0.47% in developed countries (United Nations, 2019).

Studies on UHI have been carried out primarily for areas larger than 100 km^2. However, it is noteworthy that in smaller areas, there may be alterations in the factors that influence UHI related to environmental, social and economic dimensions, similar to those in capital cities such as in the city of Baguio, Philippines, which covers an area of 57.5 km^2 (Estoque and Murayama, 2017). Therefore, there is a need to promote studies on UHI in smaller areas, as they are more likely to experience drastic changes from the effect of urban warming (D. Zhou et al., 2019), and thus in their sustainability dimensions. A higher resolution analysis could facilitate the precise identification of areas affected by this phenomenon, to establish a scheme to prioritize actions for its mitigation and contribute in the planning of sustainable cities from the micro-territory level.

The main objective of this work is to analyze from the micro-territorial level, the influence of urban growth on the UHI phenomenon over the last two decades (2000-2020). To this end, this study developed an analysis of the behavior of environmental indices and socio-economic factors, based on operations between satellite image bands and information provided by government entities. This research examines micro-territories as units of analysis within large cities, given that they facilitate the evaluation of different phenomena's behavior on a

smaller scale, as well as the selection of measures with a synergic effect that can be applied in larger areas. First, the procedure and results from this work will serve as an input for the analysis of the influence of urban growth and meteorological variables on urban sustainability. Second, it will make it possible to identify the specific contribution of micro-territories on the main components of UHI. Third, it will serve as technical support for decision-makers in the field of territorial planning.

This study is innovative in that it recognizes the importance of a bottom-up approach, developing a multiscale analysis in a territory with little vegetation, located near the equatorial zone at an altitude of 2625 meters. The combination of the analysis of satellite images and census information with statistical procedures performed, in order to identify the relationships of urban growth, UHI, and sustainability in the micro-territory further highlights the contributions and innovation of this research.

2. Methods

2.1 Study area

The locality of Kennedy was the territory selected for the case study; it is located in the southwest of Bogotá, the capital of Colombia (see Fig.1) at an altitude of 2625 meters on a high plateau on the eastern slopes of the Colombian Andes. According to the work done by Wu et al. (2019), in which the authors made use of a medium resolution image radiometer spectrum, Bogotá is one of the cities in Latin America with the highest daytime and nighttime UHI. Furthermore, according to Ramírez-Aguilar and Lucas (2019), Kennedy has the most intense UHI in the city. The study area is characterized as a space lacking in vegetation tied to an uban transformation process accentuated in the west of the locality.

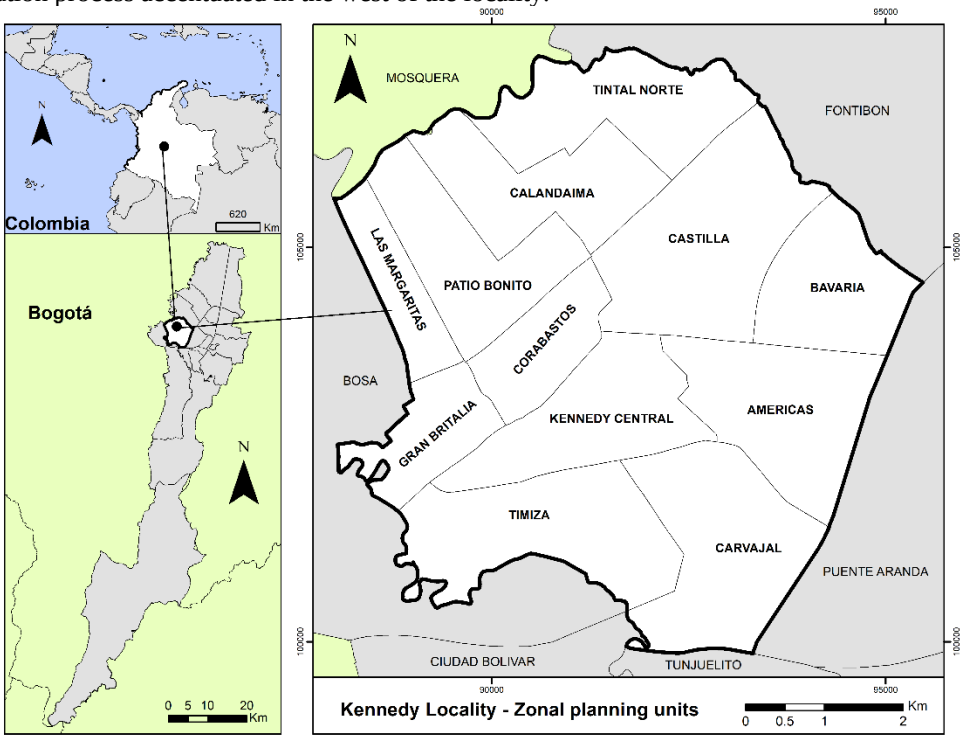


Fig. 1. Location of the study area

Kennedy is one of twenty localities in the capital city and is located in a flat area where important water sources are found, such as the Bogota, Fucha, and Tunjuelo Rivers, along with the La Vaca, El Burro, and El Techo Wetlands.

In 2018, Kennedy had 1,230,539 inhabitants, approximately 15% of the city's total population, with an average population growth rate of 2.5% per year. The locality has a total area of 38.58 km^2, of which, 93.4% is urban,

with 6.5% corresponding to urban expansion (Veeduría Distrital, 2018).

The locality is distributed into twelve zonal planning units (ZPU) (see Fig.1), which were implemented to better manage urban development planning. There are several economic activities in the locality including the city's main supply center. Kennedy has a system of public transportation portals and road infrastructure for access and transit in the city.

In 2012, according to government entities, most of the locality had an average buildability of 2 – 4 floors per block, and by 2020, the height increased to 4 – 5 floors in the ZPUs of Corabastos, Castilla, and Calandaima. The ZPUs of Bavaria and Castilla have specific points with buildings higher than 15 floors.

Meteorological data for the study area are monitored by the Carvajal Sevillana station to the south of the locality (at ZPU of Carvajal) and the Kennedy station at Kennedy Central ZPU. During the period from 2008–2019, an annual average surface temperature of 15.4°C was recorded in Carvajal Sevillana and 14.9°C in Kennedy stations, respectively (SDA, 2020).

Concerning the average accumulated precipitation, during 2008–2019, the Carvajal station recorded 725.7 mm with the Kennedy station logging 828 mm. On average, the months with the most rainfall in 2019 were May, October, and November. The first and third quarters of the year registered the lowest rainfall (SDA, 2020).
Kennedy is a city sector with environmental and socio-economic characteristics that make it relevant for an analysis of urban sustainability. It is the second most populated locality in the city and has the most significant problems in terms of air quality (Ramírez-Aguilar and Souza, 2019). During 2016-2019, the WHO recommendation (2006) for PM_{10} ($50\mu g/m^3$) was exceeded by an annual average of 220 days at Carvajal station, and 165 days at Kennedy station. The largest exceedances are historically in the first quarter of the year (SDA, 2020).

2.2 Research design

This work consists of five stages sequentially developed (see Fig.2), which are described below.

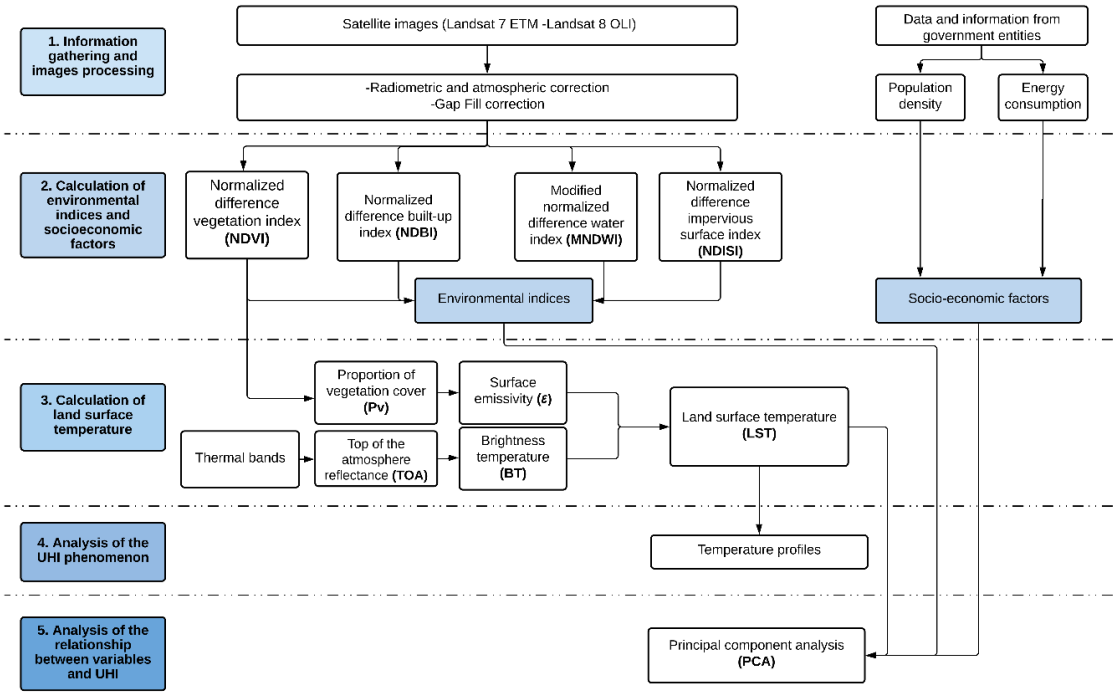


Fig. 2 Procedure to analyze the influence of urban growth on the UHI phenomenon

2.2.1 Information gathering and image processing

Several available satellite missions currently have accessible information to develop multi-temporal studies of urban phenomena, with the US Geological Survey (USGS) Landsat mission being one of the few which provide its services at no cost. Landsat is equipped with specific instruments for multispectral remote sensing and has sensors that are useful for UHI characterization and analysis (Ezimand et al., 2021).

This stage consisted of searching for and processing satellite images offered by free USGS platforms. The images were characterized by the variety of bands that they are made up of, and generally have a resolution between 15 and 30 m per pixel.

As the study period covered 2000 to 2020, Landsat 7 Enhanced Thematic Mapper (ETM+) (2000-2013) and Landsat 8 Operational Land Imager (OLI) (2013-2020) images were used. Through a year by year search, satellite images were selected from December to March, corresponding to the dry season. Furthermore, images with high cloudiness in the study area were discarded.

Based on the above criteria, the images that met the required conditions were selected (see Table 1). The climatic phenomena present each year were considered since they could have influenced UHI intensity. The El Niño phenomenon can generate temperature increases in contrast to the La Niña phenomenon. The "Neutral" condition in Table 1 indicates that neither of the two climatic phenomena occurred.

Table 1. Satellites used and atmospheric phenomenon for each year studied

Year	Month	Day	Atmospheric phenomenon	Satellite	Source
2000	Feb	20	Niña		
2002	Feb	25	Neutral		
2003	Jan	27	Niño		https://eos.com/landviewer/?lat
2004	Feb	15	Neutral	Landsat 7 ETM +	=4.64930&lng=-
2007	Feb	7	Neutral		74.06170&z=11&datasets=2
2009	Dec	29	Niño		
2012	Feb	21	Niña		
2014	Feb	2	Neutral		
2015	Feb	21	Niño	Landsat 8 OLI	https://search.remotepixel.ca/#
2018	March	17	Niña		3/16.69/-48.33
2020	March	22	Niño		

Given the local atmospheric conditions, lighting, and cloudiness present during data acquisition, the images were subjected to radiometric and atmospheric correction. As such, radiometric correction was applied to the images via the FLAASH method, using the ENVI Program Version 5.3, to manipulate the pixel values and obtain the most homogeneous intensity values, and even correct errors in the pixels. In a complementary manner, an atmospheric correction was performed to reduce the effect of aerosols, as well as the radiance introduced to the sensor reflected in the image (Aguilar et al., 2014). Due to sensor failure causing information losses in certain sections of the images, a gap fill correction was carried out on different Landsat 7 EMT+ images (2004-2007-2009-2012) through a simple triangulation method.

Socio-economic factors were addressed given the deterioration of natural resources. However, due to limitations regarding access to information for micro-territories such as the study area, it was not possible to establish socio-economic information at the spatial level with the characteristics of the environmental indices. To overcome these limitations, census data and information published by government entities were reviewed. In this manner, it was possible to demonstrate changes by ZPU over the years with respect to population density and energy consumption. These two factors influence urban growth processes, which put pressure on environmental components and urbanization.

2.2.2 Calculation of environmental indices

Once the satellite images were processed, the spectral indices were calculated through operations with the images' bands. The normalized difference vegetation index (NDVI), the normalized difference built-up index

(NDBI), and the normalized difference water index (NDWI) have been used in UHI (Chen et al., 2006; Grigoraș and Urițescu, 2019; Kikon et al., 2016; Min et al., 2019), and have demonstrated a significant correlation with the (LST) that functions as a parameter to control the water and energy balance between the atmosphere and the land surface. Table 2 describes the equations used to calculate the spectral indices.

Table 2. Biophysical indices for environmental factors

Index	Equation	References
Normalized vegetation difference index enables an estimation of the quantity and quality of vegetation based on the portion of red light absorbed and the near infrared reflected. The index ranges from -1 to 1, in which negative values correspond to water surfaces, rocks, or artificial structures and positive values represent vegetation.	$\text{NDVI} = \frac{NIR - RED}{NIR + RED}$ (1)	(Grigoraș and Urițescu, 2019; Madanian et al., 2018; Yuan and Bauer, 2007)
Normalized built-up difference index enables the identification and estimate of built or under construction areas. It also facilitates the analysis of urban growth and built-up areas. The index values range from -1 to 1; negative results indicate the presence of vegetation, and positive values correspond to built-up areas or anthropogenic infrastructures.	$\text{NDBI} = \frac{SWIR - NIR}{SWIR + NIR}$ (2)	(Musse et al., 2018; Zha et al., 2003)
Modified normalized difference water index enables the recognition of water covers, isolating them from other coverings. Its range is -1 to 1; the positive values are interpreted as water; and values close to zero or negative indicate vegetation or soil.	$\text{MNDWI} = \frac{GREEN - SWIR}{GREEN + SWIR}$ (4)	(Chen and Zhang, 2017; Xu, 2006)
Normalized difference impervious surface index: this index has been used to extract impervious surfaces. NDISI removes noise such as soil and water. The surface radiation is maximized by using the thermal wavelength (TIR), minimizing the reflectance of NIR, SWIR, and GREEN per impermeable surface. Positive values represent impermeable surfaces, as opposed to negative values.	$\text{NDISI} = \frac{TIR - \frac{(GREEN+NIR+SWIR)}{3}}{TIR + \frac{(GREEN+NIR+SWIR)}{3}}$ (5)	(Estoque and Murayama, 2015; Musse et al., 2018; Xu, 2010)

In which, NIR is near infrared, RED is the red band, SWIR is short-wave infrared 1, which differentiates soil and vegetation moisture; this wave penetrates through thin clouds. GREEN corresponds to the green band. For Landsat 7, these bands are 4, 3, 5 and 2; and for Landsat 8, they are 5, 4, 6 and 3.

2.2.3 Calculating Land Surface Temperature (LST)

The images' thermal bands, band 6 for Landsat 7 and band 10 for Landsat 8, were used to calculate LST. Using the method proposed by USGS (Ihlen and USGS, 2019a, 2019b) and through equations (6) and (7), a conversion of the digital number to a radiometric scale was performed.

For Landsat 7,

$$L\lambda = \left(\frac{LMAX_\lambda - LMIN_\lambda}{QCALMAX - QCALMIN}\right) * (QCAL - QCALMIN) + LMIN_\lambda \quad (6)$$

In which Lλ is the reflectance of the top of the atmosphere (TOA) in $\frac{W}{m^2 * sr * \mu m}$; $LMAX_\lambda$ and $LMIN_\lambda$ are radiance values obtained from image metadata; QCAL is the quantified pixel value calibrated in a digital number; QCALMAX and QCALMIN are the maximum and minimum pixel of band 6. The images' digital numbers were transformed into radiation units.
For Landsat 8,

$$L\lambda = M_L * QCAL + A_L \, (7)$$

In which Lλ is TOA in $\frac{W}{m^2 * sr * \mu m}$; M_L is the multiplicative brightness scale factor for band 10; A_L is the additive radiance scale factor for the same band; and QCAL is the quantified value of the digitally calibrated pixel.

The brightness temperature (TB) was then calculated using equation (8), which enables the irradiation to be transformed into surface temperature in degrees Kelvin (Ihlen and USGS, 2019b, 2019a).

$$TB = \frac{K_2}{\ln\left(\frac{K_1}{L\lambda} + 1\right)} \, (8)$$

In which K_1 and K_2 are calibration constants taken from the image metadata. Lastly, the LST is calculated via equation (9); the results are presented in degrees Kelvin.

$$LST = \frac{TB}{1 + \left[\lambda * \frac{TB}{a}\right) \ln \varepsilon} \, (9)$$

In which λ is the wavelength of the radiance emitted; a is 1.438×10^{-2} mK (Estoque and Murayama, 2017; Senanayake et al., 2013) and ε is the surface emissivity, which is calculated by equation (10) (Grigoraș and Urițescu, 2019; Wang et al., 2018).

$$\varepsilon = 0.004 * Pv + 0.986 \, (10)$$

In which Pv is the vegetation proportion calculated as shown in equation below (11):

$$Pv = \left[\frac{(NDVI - NDVI_{min})}{(NDVI_{max} - NDVI_{min})}\right]^2 (11)$$

2.2.4 Analysis of UHI behavior

To understand UHI behavior, distance vs. temperature profiles were made in four different directions: 1) north-south, 2) northwest-southwest, 3) west-east, and 4) northeast-southwest; with the pixel value determined every 500 meters. The year by year results were categorized by the dominant climate phenomenon (El Niño or La Niña in each case) and based on these profiles, the areas with the highest or lowest temperatures in the locality were identified.

2.2.5 Association analysis of the variables

Urban sustainability is primarily related to the behavior of the environmental, social, and economic dimensions; in which urban growth generates a series of pressure points that can be seen in these dimensions' behavior.
A principal component analysis (PCA) was performed to identify the degree of relationship between variables. The PCA also makes it possible to identify the contribution of variables (environmental and socio-economic factors) and individuals (each point in the micro-territory) to the main components.

The input information consisted of a band composition from the raster images of the spectral indices presented in Table 2, along with the population density and energy consumption variables. The band composition was performed with the ArcGIS software 10.8.1, followed by the PCA analysis carried out with the free access software R.

3. Results and discussion

3.1 Environmental factors

Rapid urbanization has affected the natural environment of urban areas. Consequently, the UHI phenomenon occurs, which are created and intensified by the increase of impermeable surfaces or heat produced by human activities, as well as by the reduction of green spaces in a territory. In this vein, vegetation dynamics, the expansion of built-upon soil, and water bodies are environmental factors that can influence the formation and behavior of UHI.

3.1.1 Vegetation dynamics

Vegetation dynamics were analyzed based on the NDVI from 2000 to 2020 (see Fig.3). Most of the green area was located to the north of the locality. In 2000, about 36% of the study area (13.98 km^2) corresponded to a zone with vegetation. A notable reduction in vegetation has occurred since 2003, primarily attributed to the increase in building construction in the area.

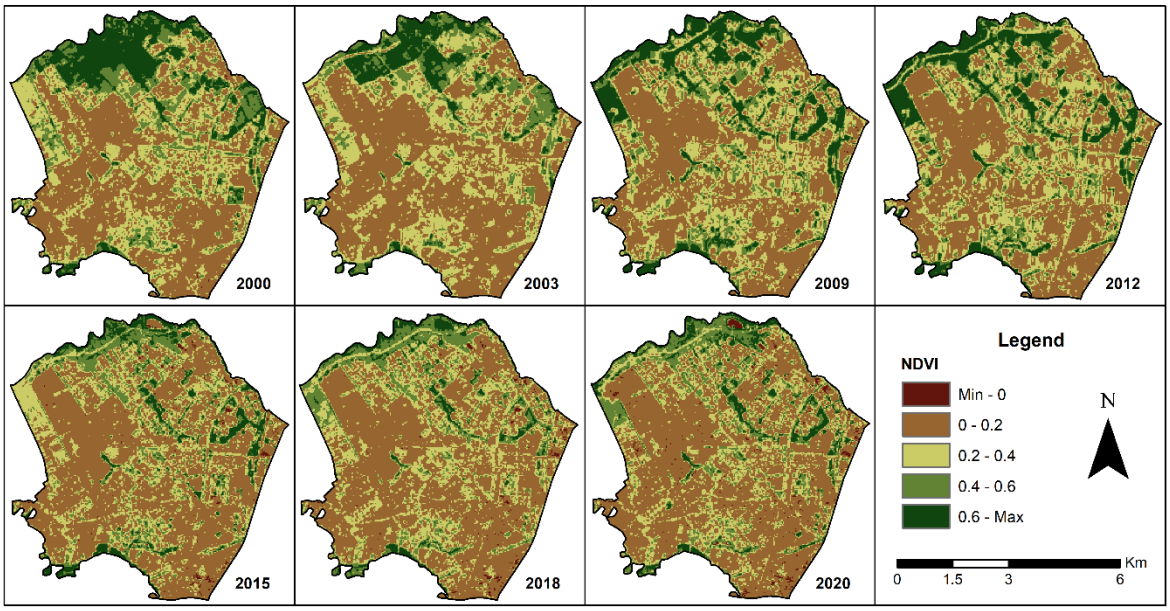


Fig. 3 Dynamics of the Normalized Difference Vegetation Index

By 2009, vegetation had been reduced by 9.62 km^2, and the first consolidation of buildings in the northern zone can be detected. In 2020, there are approximately 7.14 km^2 of vegetation areas, which fall in the NDVI range of 0.2 to values > 0.6 (see Fig.3). The range between 0.2 and 0.4 corresponds to areas with scarce or dispersed vegetation; between 0.4 and 0.6 corresponds to areas with moderate vegetation; and NDVI values greater than 0.6 represent locations where the density of vegetation is most likely green and healthy.

3.1.2 Built-up areas and impervious surfaces

Kennedy is mostly covered by buildings and impermeable surfaces such as roads and sidewalks with scarce vegetation. The behavior of bare soils or built-up covers is inverse to that of the vegetation (see Fig.4). In 2000, the area with buildings was approximately 22.1 km^2, which were consolidated in the southern part of the locality. Seven years later, the area with buildings increased to 26.9 km^2, reaching 28.26 km^2 in 2020. In the first years of the study, land occupancy for housing in illegal urbanizations continued in areas such as the La Vaca Wetland (in the Corabastos ZPU), in the northern part of the locality in areas of El Tintal, particularly on the banks of the

Bogotá River, northwest of Kennedy (Escobar Franco, 2012). The greatest variations occurred in the northern part of the locality.

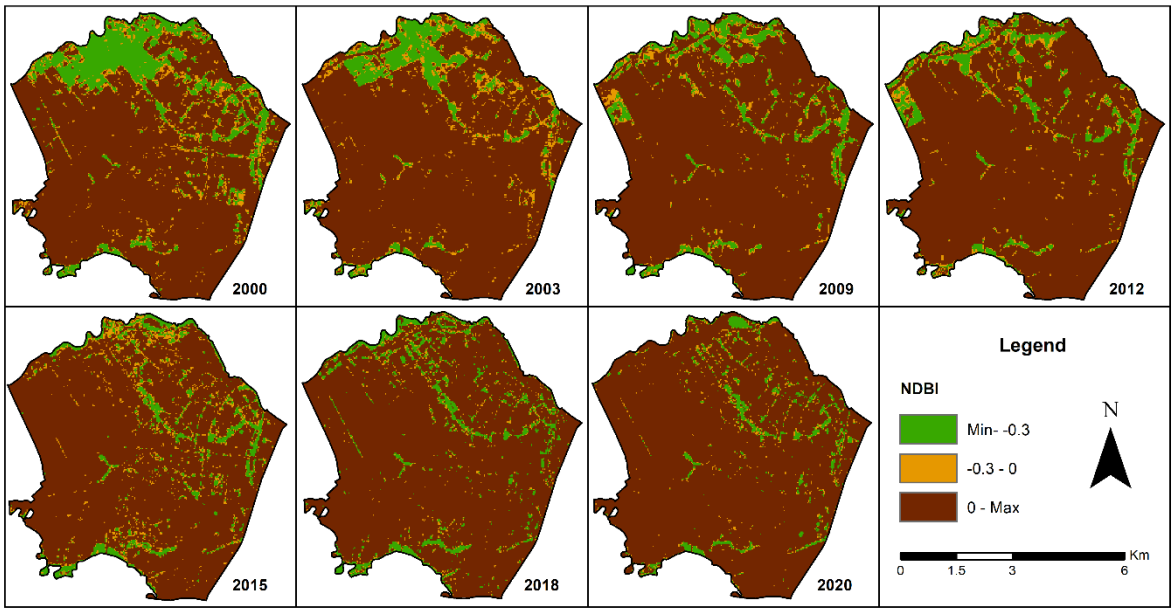


Fig. 4 Dynamics of the Normalized Differential Build-up Index

3.1.3 Bodies of water

The bodies of water were evaluated by measuring the MNDWI. Figure 5 shows the years in which changes occurred. In 2000, Lake Timiza, which previously could not be seen, stands out in the southern area. In the following years, there were no variations in the MNDWI. However, based on the spatial operations with geographic information of the city, a reduction of the water mirrors was identified, from 0.032km^2 in 2000 to 0.0154 km^2 in 2014. This reduction occurred in the Wetlands located in the center and north of the locality.

It is possible, that the reduction in vegetation recorded over the years may have revealed the bodies of water. Grasslands also reduced from 0.895 km^2 in 2000 to 0.263 km^2 in 2014. For this reason, as of 2014, the Pondaje lagoon, created to regulate the flow of water and prevent flooding in the area, is seen in the northern part of the locality. By 2020, the El Burro and La Vaca Wetlands in the center of the locality can be seen to a lesser extent.

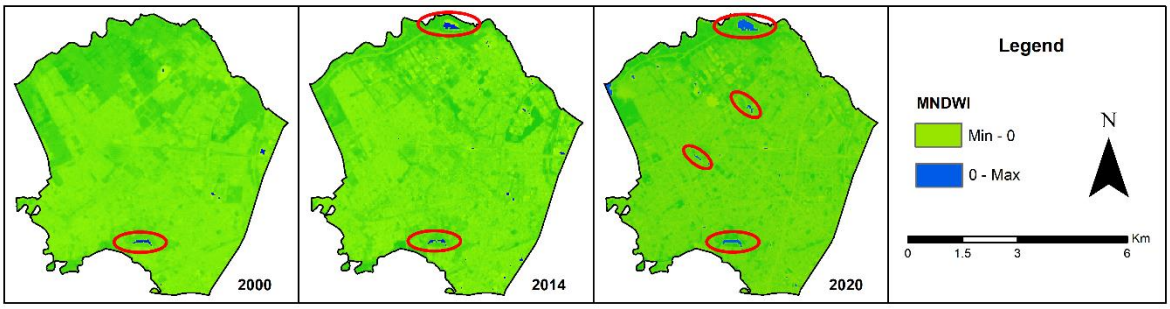


Fig. 5 Bodies of water

3.2 Socio-economic factors

The socio-economic factors analyzed correspond to changes in population density and energy consumption flows, which are integral components of the development of urban spaces. The larger the population, the greater the pressure on resources, and the greater the energy requirements. Sustainable cities and communities entail balancing pressures generated to guarantee well-being and quality of life, as they are committed to providing adequate housing, access to transport systems, increased inclusive and sustainable urbanization, in addition to safeguarding its natural heritage, reducting of environmental impacts, and universal access to green areas (UNDP, 2020). These challenges are intensifying for urban areas such as Kennedy. The dynamics of local population density and energy consumption are analyzed below.

3.2.1 Population density

Population density maps were created based on census data (SDP, 2020) (see Fig.6). In 2005, the average population density of the locality was 24,625 inhabitants/km^2, ranging from 3,800 to 50,500 inhabitants/km^2. Patio Bonito was the most densely populated ZPU in the city, exceeding the gross density of Bogotá. In recent years, population density figures have exceeded those in other cities in Latin America such as Quito, Ecuador (5401inhabitants/km^2) and Mexico City, Mexico (5966 inhabitants/km^2).

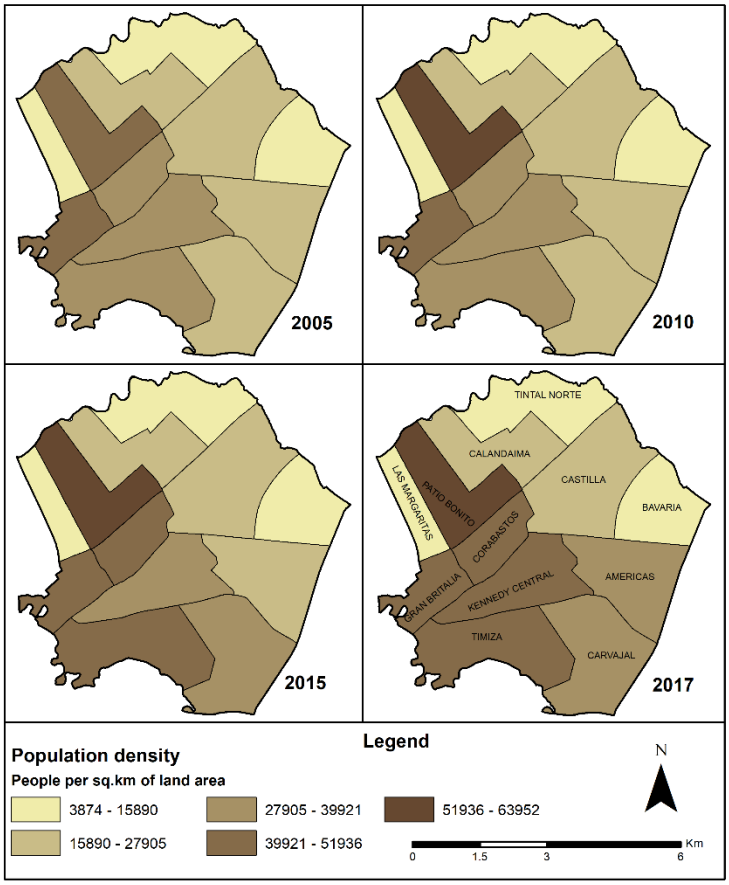


Fig. 6 Population density by ZPU in Kennedy

An increase in population density over the years can be seen in the central zone and south of the locality in the ZPUs of Corabastos, Kennedy Central, Timiza, Carvajal, and Américas.

3.2.2 Energy consumption

Kennedy is one of the localities of Bogota with the highest concentrations of electric energy consumption (Alcaldía Mayor de Bogotá, 2017). Over the years, there has been a progressive increase in the consumption of energy for residential, commercial, and industrial use (see Fig.7). However, in the eastern part of the locality, industrial consumption has decreased, while residential and commercial consumption have increased. The ZPUs

of Patio Bonito, Timiza, and Castilla are the areas with the highest residential energy consumption; Carvajal has the highest consumption for commercial and industrial use.

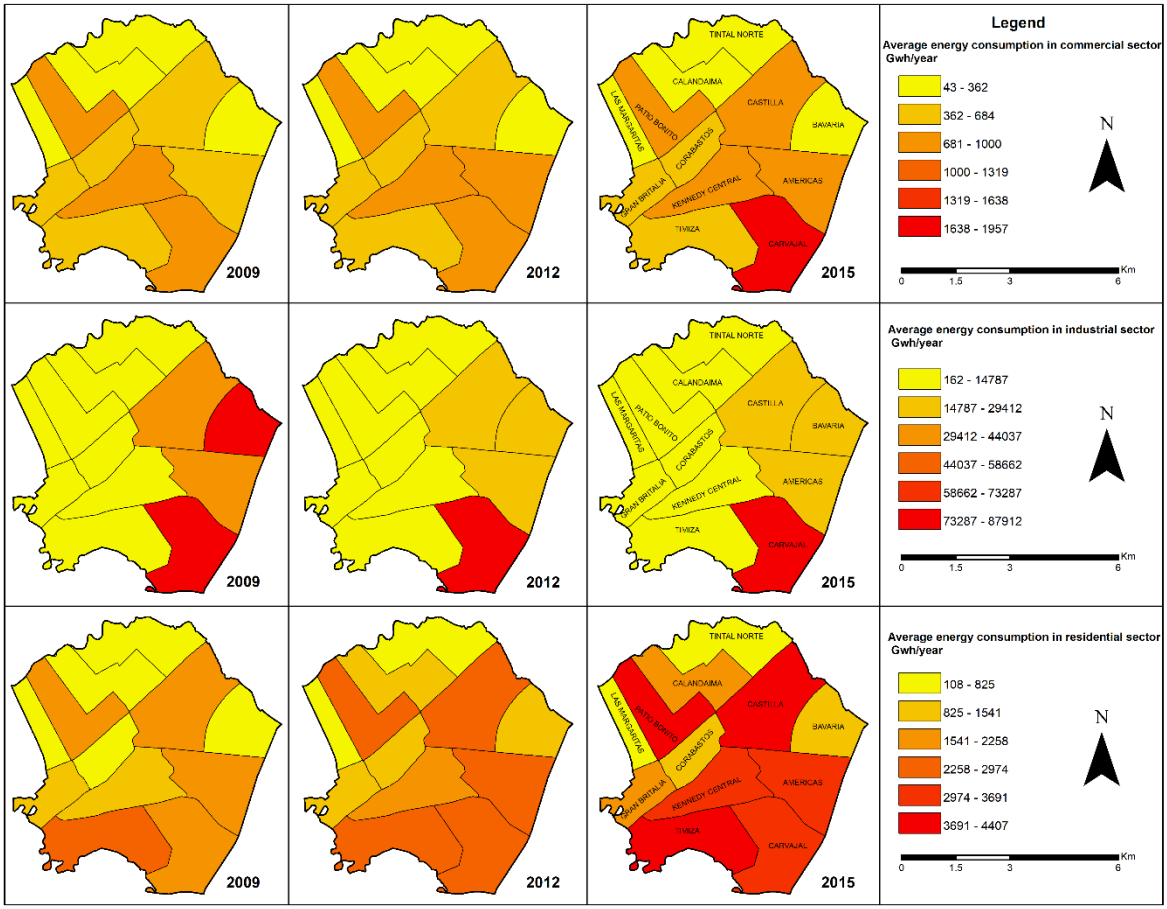


Fig. 7 Dynamic of the energy consumption in Kennedy

3.3 LST spatial-temporal pattern

The LST distribution was classified in ranges of 3ºC (see Fig.8). The lowest temperatures occurred in 2000 in the north of the locality due to the presence of healthy consolidated vegetation. Two years later, temperatures increased in the eastern and southern parts of the locality, ranging from 28° to 33ºC. The LST distribution was more uniform throughout the area with values between 12° and 33°C during the following year. However, small areas in the center of the locality stand out, such as the ZPUs of Corabastos and Kennedy Central, where temperatures are higher than 28ºC. This pattern is seen in every year, and from 2012 its increase exceeds 5° Celsius.

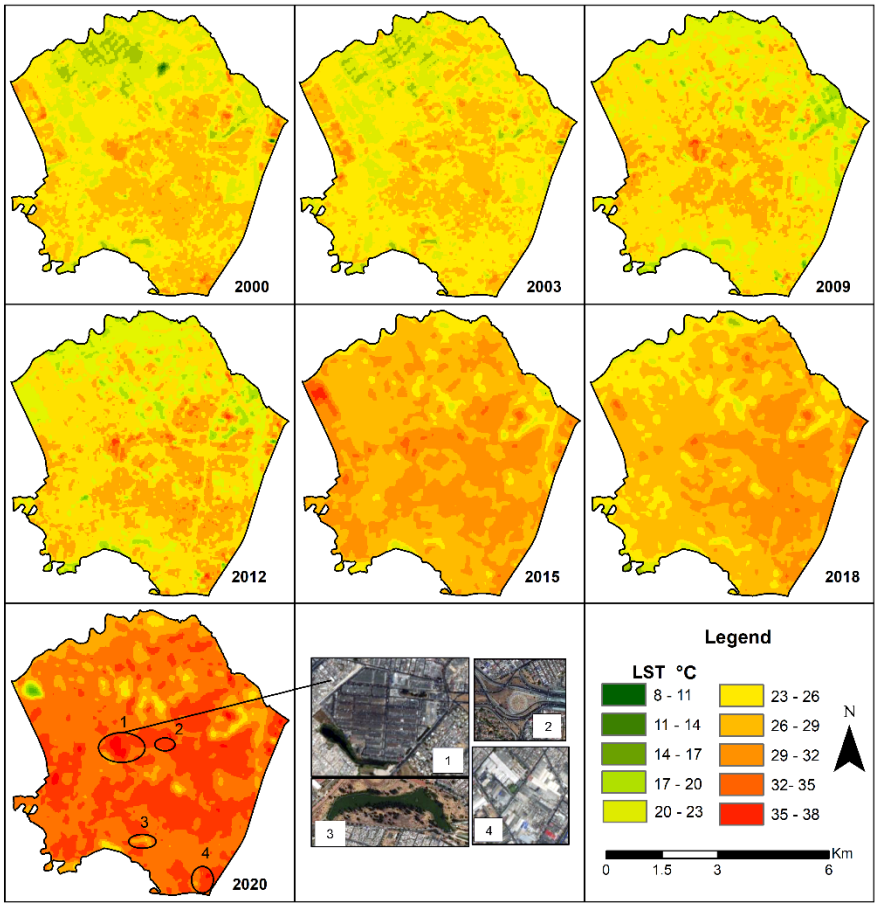


Fig. 8 Land surface temperature dynamic

These results were contrasted with the locality's economic and urban conditions. The area with the highest temperatures coincides with the location of the country's main supply center; Corabastos (image 1 in Figure 8). Approximately 1,200 vehicles carrying supplies enter the area every day, most of which are older models, which emit atmospheric pollutants. Temperature increases cause an accelerated production of smog, concentration of pollutants, and impacts on local meteorology (Ngarambe et al., 2021), which cause PM_{10} to exceed WHO recommendations (WHO, 2006). The increase in residential and commercial energy consumption in the south and east of the locality also contributes to this process.

Since 2014, there has been a homogeneity of temperature distribution changes, with approximately half of the locality having temperatures between 26º and 33º Celsius in the same south-east area. In 2015, temperatures intensified in most of the territory due to the presence of the El Niño phenomenon, while in 2018, temperatures decreased, which can be attributed to the precipitation generated by the La Niña phenomenon. In 2020, the LST was greater than 26ºC throughout most of the locality, with a maximum of 38ºC, which also coincide with the ZPUs with high energy consumption for the different analyzed uses.

As Kennedy is a locality with low levels of vegetation (0.35 trees/km^2 in 2019) compared to built-up areas, it is vulnerable to continue experiencing the intensity of UHI. In certain micro-territories, there has been uncontrolled urbanization, mainly in peri-urban areas, which has affected land use and increased urban expansion in natural areas (Dobbs et al., 2018). This contrasts the fact that 25% of urban areas are unplanned or informally planned at the global level (UN-Habitat, 2019).

The LST behavior was analyzed by profiles (see Fig.9), in which the locations with the highest temperatures in the micro-territories were highlighted. The blue lines represent the years in which the La Niña phenomenon occurred, the red lines represent the El Niño phenomenon, and the green lines indicate the neutral years. Figure

9 with a north-south heading, shows the extreme low temperature values, and a peak in the center of the locality, where a mass public transport station is located (Banderas station, image 2 in Figure 8). At this site, the highest temperatures were reached, surpassing 33ºC in 2018.

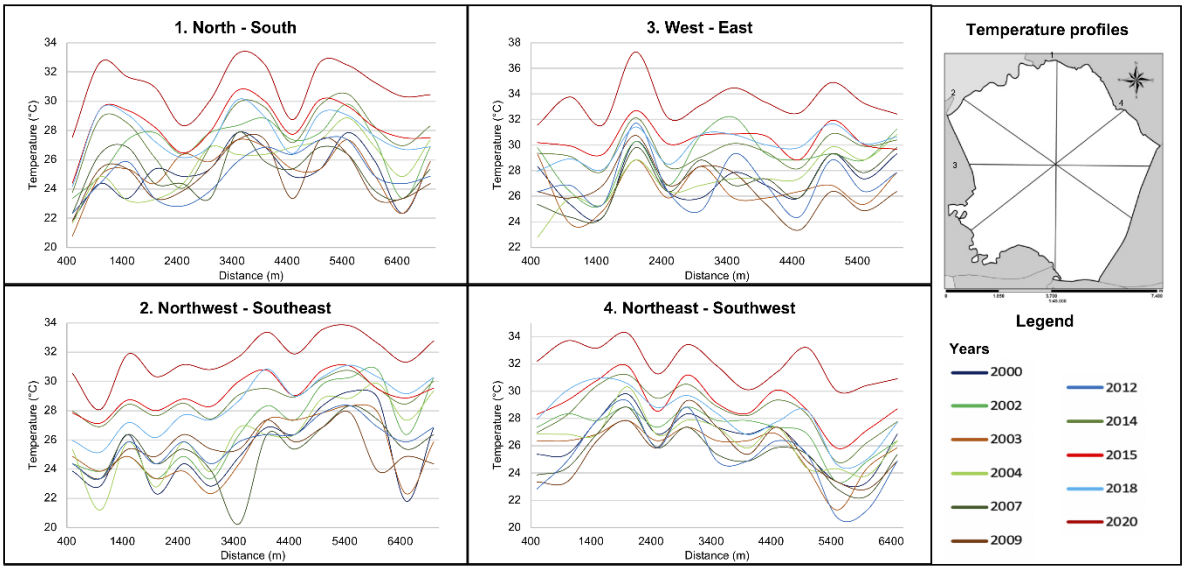


Fig. 9 Land surface temperature profiles

Despite most of the curves showing a uniform behavior in years in which the El Niño phenomenon was present (see the red curves in Fig.9), the behavior is above the others. The graphs show a similar behavior between 2015 and 2020, but in 2020 the temperature increased by approximately 2ºC when compared to 2015.

In the west-east direction, there is a temperature increase in the first 1500 meters, coinciding with the central supply center which consists of a consolidated built area of 420,000 m^2. This area is known for its economic activity and daily vehicle movement. Moreover, the average building height is 5 – 6 floors per block in this location, which can hinder proper air circulation.

It is important to mention that there is a noteworthy pattern of temperature decreases linked to bodies of water. Nevertheless, in the 20 years analyzed, the temperature rose by approximately 3.6º an annually progressive increase. This rise means that minimum temperatures are mostly above 14ºC and maximum temperatures average 34.7º, with the highest temperature in 2020 being 37.84º Celsius. There was an increase in the years when El Niño occurred; however, it did not change the trends in LST behavior.

Studies in the global context mostly analyze areas larger than 100 km^2 and none have focused on local areas, or micro-territories. When analyzing UHI behavior by profile, as was the case in the study developed by Estoque and Murayama (2017), the UHI pattern in Kennedy largely held steady during the last years of study from 2015 to 2020. The first years of the study had low temperature values, as there was a greater presence of vegetation (average temperature of 22.5ºC during 2000-2003).

3.4 Relations between UHI and impact factors

The loss of vegetation coverage has resulted in an increase of UHI in the locality, as the amount of vegetation influences the LST by the heat flow from the surface through evapotranspiration. Furthermore, trees provide shade and cooling that can prevent direct exposure of land surfaces to solar radiation (Singh et al., 2017; Soltani and Sharifi, 2017).

The PCA analysis made it possible to establish three components that account for 75.3% of the variance in the data. The first component: Dim 1 (37.7% of the variance), highlights central elements of urban expansion. The variables with the greatest contribution (67%) in this component were NDVI, BU, NDBI, and MNDWI, with correlations greater than 70% with Dim 1. Although weak, there is also a positive correlation with population density (55%), residential energy consumption (53%), industrial energy consumption (50%), and LST (49%), which are elements that characterize the effects of urban growth. As in the study developed by Chen and Zhang (2017), the relationship between LST and NDBI had one of the strongest linear positive correlations, which can be attributed to the heterogeneity of the land surface, particularly in areas with a low fraction of vegetation. The relationship between LST, and NDBI is linearly positive, given that when built-up areas or soils without vegetation increase, temperature is not absorbed or regulated, which generates an increase of temperatures in urban centers.

The second component Dim 2, correlates energy consumption in the different analyzed uses (19.9% of the variance) with a 56.6% contribution to this component. The relationship between the first and second components is shown in Fig. 10a, where both the quality of the variables and the correlation with the components can be seen.

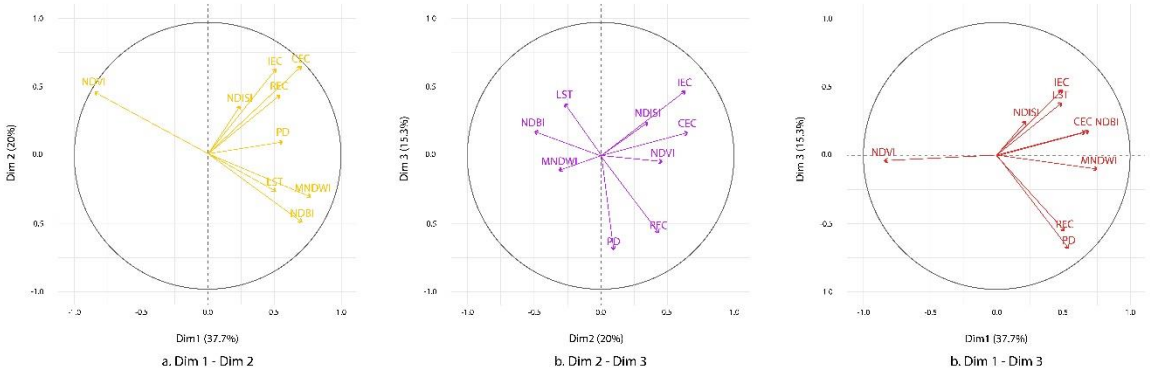


Fig. 10 Principal component analysis

In the clusters, it is also noted that LST and NDVI have a moderate negative correlation (-0.67), given that lower temperatures occur in areas with dense vegetation. NDVI also has an inverse correlation with NDBI (-0.95); greater urban growth and the decrease in vegetation are correlated with UHI intensity.

The third component Dim 3, called LST, accounts for 15.3% of the variance in the data and includes population density variables, energy consumption by residential and industrial users, the NDISI, and LST; the variables that contribute the most to the component (94.2%). Figures 10b and 10c show the relationship of this component with Dim 1 and 2. In each case, the correlation of the variables with each component is evident by the fact that they are close to the edge of the circumference.

These analyses can also be used to compare dimensions and identify the micro-territories that contribute the most to the components (see Fig.11). As such, the largest contribution were found in the ZPUs of Patio Bonito, Carvajal, and Tintal Norte, which coincide with the areas with the highest population density, high energy consumption, growth in built-up areas, and green areas reduction.

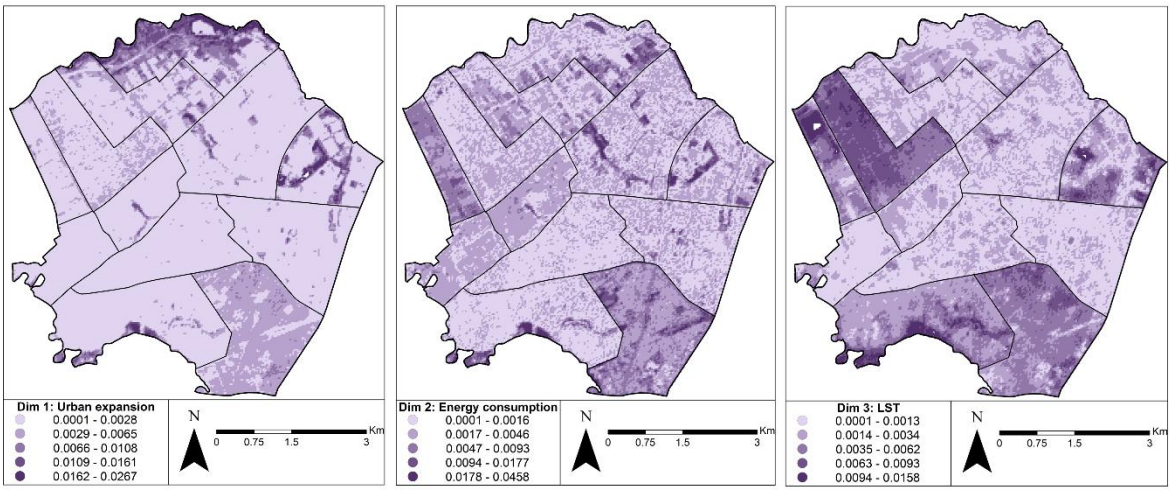


Fig. 11 Contribution from micro-territories to the components

Figure 11 highlights the contribution from the north and east of the locality to urban expansion. North and south of Kennedy contribute more to energy consumption (Dim 2). The territories that contribute to component 3 (see Fig.11 Dim 3: LST) are mainly in the south and west of Kennedy. A mix of commercial, residential, and industrial activities is characteristic in those areas (Patio Bonito, Timiza, and Carvajal ZPU's). The above once again demonstrates the implications of urban growth on urban sustainability conditions.

4. Discussion

4.1 Influence of urban growth on the UHI phenomenon

The formation of UHI occurs mainly in the center of the locality, as it is the area with the highest vehicle mobility and consolidated residential and commercial areas. These findings are similar to other studies, such as the one developed by Amanollahi et al. (2016), which presented the critical points of UHI in parts of the city with commercial and residential areas, main roads, and even in areas for agricultural use. In Kennedy, LST progressively increased, with a notable homogenization of the temperature over the entire locality for the last years of the study. This situation shows the relevance in extending the resolution of spatial analysis, since the behavior of UHI reflected by the ZPU in Kennedy as a territory within a large city behaves similarly in capital cities.

The case in the study developed for Bogotá by Ramírez-Aguilar and Lucas (2019) demonstrated the relationship between population density and heat intensity. Moreover, according to (Zhang et al., 2017), there is an indirect relationship between those factors, since population density is a driver of different land uses and economic activities.

In this study on the micro-territory of Kennedy, activities related to population density were added, which include greater vehicle traffic entering urbanized areas and the operation of commercial areas with their corresponding energy consumption.

Territories with high levels of air pollutants, as is the case of Kennedy, coincide with the presence of different economic activities, high population density, increased impervious surfaces, and urban conditions that could experience the UHI phenomenon. The above is comparable with the results found by Bokaie et al., (2016), and is reflected in this study in the analysis of environmental and socio-economic indices, as well as the PCA.

Changes in land cover play an important role in the development of UHI, as vegetation is replaced by impermeable surfaces and the deterioration of environmental quality increases. It is important to consider population growth trends, as population increases result in a greater demand for resources such as water, energy, and soil. As these areas do not have new zones for construction, low buildings will inevitably be replaced by high buildings to house a larger population, modifying the morphology of the land, which therefore causes

changes in air quality and temperature. This situation will require measures to be adopted to balance temperatures; for example, green belts in different areas, particularly those that contribute the most to components 1 and 3.

The procedure developed in this research study can be applied to several urban areas to identify the territories that contribute the most to UHI, in addition to the most appropriate urban and landscape planning measures. It is also applicable to cities such as Ghaziabad (India), one of the most polluted cities in the world with a population density comparable to some of the ZPUs in Kennedy, as well as Orangi Town in Karachi (Pakistan) and Neza (Mexico), which are among the world's largest suburbs in cities with the highest air pollutant records.

4.2 Implications of UHI on urban sustainability

Urban areas face major challenges in terms of sustainability, as they must balance the demand for resources inherent to urban growth with existing ecosystems. Consequently, it is necessary to not only establish measures to mitigate local pollutants from mobile and stationary sources, but to also create sustainable micro-territories including buffer zones for environmental aspects. Future research should correspond to establishing measures and analyzing correlation with reducing the causes of UHI in each micro-territory analyzed in this study.

The approach outlined in this research study contributes to defining specific measures regarding landscape design and its potential to mitigate UHI and local pollutants. Given the difficulty of creating green areas in densely populated areas, one way to mitigate the effects of UHI in Kennedy is to improve its vegetation cover, either on roofs and green walls, or by restructuring buildings to increase the number of trees in the area. Other measures include adopting energy efficiency policies to reduce unintentional heat-generating emissions in urban areas, which can contribute at the micro-territorial level, in addition to implementing measures at a larger scale. For their part, urban and landscape planning processes require using new elements, such as different materials in buildings and infrastructures that reflect radiation and enable an LST balance to be maintained.

5. Conclusions

In comparison with studies carried out on cities and countries, research that performs higher spatial resolution analyses are more able to clearly identify the specific causes of temperature increases in a micro-territory. These studies provide support in the formation of mitigation and adaptation measures to develop sustainable cities and communities.

Furthermore, by establishing through PCA, the three components that account for the variability of the data (75%) and the subsequent identification of the micro-territories (Patio Bonito, Carvajal, and Tintal Norte) that contributed the most to the components; urban expansion, energy consumption and LST, it is possible to recognize the causes of the UHI phenomenon and its location in the micro-territory.

The spatial-temporal variation of UHI in the first years of the study shows a homogeneous temperature behavior, followed by an increase starting in 2012 in the center of the locality. The highest temperatures in the locality are reflected where there is consolidated construction, exceeding 35ºC, specifically in the Corabastos ZPU. Vehicle traffic conditions in the area influence environmental sustainability in terms of fossil fuel use, air pollutant emissions, and the development of UHI. In this study, the highest temperatures were reflected in places where vehicle traffic entails a combination of public passenger transportation and cargo vehicles. A study of UHI growth is relevant because it ties together all kinds of variables, including environmental, social and economic dimensions, which directly affect people's health and the environment.

To evaluate socio-economic factors, the population density was obtained from 2005 – 2017 census data. Patio Bonito stood out as the most densely populated ZPU while the Corabastos, Timiza, Kennedy Central, and Américas ZPUs had relevant changes when compared to 2005. It should be noted that this increase is also reflected in rises in residential energy consumption in areas where there was no substantial increase in population density, in addition to commercial activities that contributed to increases in energy consumption. The rapid

urbanization process in the north of Kennedy since 2004, reflected in a 20% increase in the built area, led to a decrease of the area with vegetation, from 13.98 km^2 in 2000 to approximately 7.14 km^2 in 2020.

This study reflects the importance of implementing mitigation strategies to reduce LTS, due to its rising trend as shown herein. This research study established the procedural approach applicable to tropical micro-territories, the results and analysis of which are comparable with other areas where progress is being made in organizing urban areas. Using this established procedure is a tool to monitor challenges related to sustainable development goals, primarily concerning sustainable cities and communities.

Declarations

Authors' contributions: All authors contributed to the conception and design of the study. Data collection, analysis and interpretation were performed by Nidia Isabel Molina-Gómez, Laura Marcela Varon-Bravo and Ronal Sierra-Parada. Nidia Isabel Molina-Gómez and Laura Marcela Varon-Bravo wrote the original draft; Nidia Isabel Molina-Gómez, Laura Marcela Varon-Bravo, Ronal Sierra-Parada and P. Amparo López-Jiménez wrote, reviewed and edited the final manuscript; and P. Amparo López-Jiménez was involved in supervision. All authors have read and approved the final manuscript.

Funding: No funding was received to assist with the preparation of this manuscript.

Conflict of interest/Competing interests: The authors declare that they have no conflict of interest.

References

Aguilar, H., Mora, R., & Vargas, C. (2014). Atmospheric Correction Methodology for Aster, Rapideye, Spot 2 and Landsat 8 Images with Envi Flaash Module Software. Revista Geográfica de América Central, 2(53), 39–59. https://doi.org/http://dx.doi.org/10.15359/rgac.2-53.2

Alcaldía Mayor de Bogotá. (2017). Consumos energéticos urbanos por usos y actividades económicas por UPZ en Bogotá DC 2009-2012-2015. Bogotá.

Amanollahi, J., Tzanis, C., Ramli, M. F., & Abdullah, A. M. (2016). Urban heat evolution in a tropical area utilizing Landsat imagery. Atmospheric Research, 167, 175–182. https://doi.org/10.1016/j.atmosres.2015.07.019

Bokaie, M., Zarkesh, M. K., Arasteh, P. D., & Hosseini, A. (2016). Assessment of Urban Heat Island based on the relationship between land surface temperature and Land Use/ Land Cover in Tehran. Sustainable Cities and Society, 23, 94–104. https://doi.org/10.1016/j.scs.2016.03.009

Carpio, M., González, Á., González, M., & Verichev, K. (2020). Influence of pavements on the urban heat island phenomenon: A scientific evolution analysis. Energy and Buildings, 226, 110379. https://doi.org/10.1016/j.enbuild.2020.110379

Chen, X. L., Zhao, H. M., Li, P. X., & Yin, Z. Y. (2006). Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. Remote Sensing of Environment, 104(2), 133–146. https://doi.org/10.1016/j.rse.2005.11.016

Chen, X., & Zhang, Y. (2017). Impacts of urban surface characteristics on spatiotemporal pattern of land surface temperature in Kunming of China. Sustainable Cities and Society, 32, 87–99. https://doi.org/10.1016/j.scs.2017.03.013

Dobbs, C., Hernández-Moreno, Á., Reyes-Paecke, S., & Miranda, M. D. (2018). Exploring temporal dynamics of urban ecosystem services in Latin America: The case of Bogota (Colombia) and Santiago (Chile). Ecological Indicators, 85(November 2017), 1068–1080. https://doi.org/10.1016/j.ecolind.2017.11.062

Escobar Franco, L. F. (2012). Plan Ambiental Local Kennedy 2013-2016. Alcaldía Local de Kennedy, 1–68. Retrieved from http://ambientebogota.gov.co/documents/10157/2883162/PAL+Kennedy+2013-2016.pdf

Estoque, R. C., & Murayama, Y. (2017). Monitoring surface urban heat island formation in a tropical mountain city using Landsat data (1987–2015). ISPRS Journal of Photogrammetry and Remote Sensing, 133, 18–29. https://doi.org/10.1016/j.isprsjprs.2017.09.008

Ezimand, K., Chahardoli, M., Azadbakht, M., & Matkan, A. A. (2021). Spatiotemporal analysis of land surface temperature using multi-temporal and multi-sensor image fusion techniques. Sustainable Cities and Society, 64(March 2020), 102508. https://doi.org/10.1016/j.scs.2020.102508

Grigoraș, G., & Urițescu, B. (2019). Land Use/Land Cover changes dynamics and their effects on Surface Urban Heat Island in Bucharest, Romania. International Journal of Applied Earth Observation and Geoinformation, 80(March), 115–126. https://doi.org/10.1016/j.jag.2019.03.009

Ihlen, V., & USGS. (2019a). Landsat 7 (L7) Data Users Handbook (p. 151). p. 151. Retrieved from https://landsat.usgs.gov/sites/default/files/documents/LSDS-1927_L7_Data_Users_Handbook.pdf

Ihlen, V., & USGS. (2019b). Landsat 8 (L8) Data Users Handbook (p. 114). p. 114. Retrieved from https://landsat.usgs.gov/documents/Landsat8DataUsersHandbook.pdf

Kikon, N., Singh, P., Singh, S. K., & Vyas, A. (2016). Assessment of urban heat islands (UHI) of Noida City, India using multi-temporal satellite data. Sustainable Cities and Society, 22, 19–28. https://doi.org/10.1016/j.scs.2016.01.005

Litardo, J., Palme, M., Borbor-Cordova, M., Caiza, R., Macias, J., Hidalgo-Leon, R., & Soriano, G. (2020). Urban Heat Island intensity and buildings' energy needs in Duran, Ecuador: Simulation studies and proposal of mitigation strategies. Sustainable Cities and Society, 62(July), 102387. https://doi.org/10.1016/j.scs.2020.102387

Liu, X., Zhou, Y., Yue, W., Li, X., Liu, Y., & Lu, D. (2020). Spatiotemporal patterns of summer urban heat island in Beijing, China using an improved land surface temperature. Journal of Cleaner Production, 257, 120529. https://doi.org/10.1016/j.jclepro.2020.120529

Madanian, M., Soffianian, A. R., Soltani Koupai, S., Pourmanafi, S., & Momeni, M. (2018). The study of thermal pattern changes using Landsat-derived land surface temperature in the central part of Isfahan province. Sustainable Cities and Society, 39(November 2017), 650–661. https://doi.org/10.1016/j.scs.2018.03.018

Min, M., Lin, C., Duan, X., Jin, Z., & Zhang, L. (2019). Spatial distribution and driving force analysis of urban heat island effect based on raster data: A case study of the Nanjing metropolitan area, China. Sustainable Cities and Society, 50(December 2018), 101637. https://doi.org/10.1016/j.scs.2019.101637

Musse, M. A., Barona, D. A., & Santana Rodriguez, L. M. (2018). Urban environmental quality assessment using remote sensing and census data. International Journal of Applied Earth Observation and Geoinformation, 71, 95–108. https://doi.org/10.1016/j.jag.2018.05.010

Ngarambe, J., Joen, S. J., Han, C. H., & Yun, G. Y. (2021). Exploring the relationship between particulate matter, CO, SO2, NO2, O3 and urban heat island in Seoul, Korea. Journal of Hazardous Materials, 403(2), 123615. https://doi.org/10.1016/j.jhazmat.2020.123615

Oke, T. R. (1982). The energetic basis of the urban heat island (Symons Memorial Lecture, 20 May 1980). Quarterly Journal, Royal Meteorological Society, 108(455), 1–24.

Peres, L. de F., Lucena, A. J. de, Rotunno Filho, O. C., & França, J. R. de A. (2018). The urban heat island in Rio de Janeiro, Brazil, in the last 30 years using remote sensing data. International Journal of Applied Earth Observation and Geoinformation, 64(October 2016), 104–116. https://doi.org/10.1016/j.jag.2017.08.012

Portela, C. I., Massi, K. G., Rodrigues, T., & Alcântara, E. (2020). Impact of urban and industrial features on land surface temperature: Evidences from satellite thermal indices. Sustainable Cities and Society, 56(February), 102100. https://doi.org/10.1016/j.scs.2020.102100

Ramírez-Aguilar, E. A., & Lucas Souza, L. C. (2019). Urban form and population density: Influences on Urban Heat Island intensities in Bogotá, Colombia. Urban Climate, 29(May), 100497. https://doi.org/10.1016/j.uclim.2019.100497

SDA. (2020). Informe Anual de Calidad del aire de Bogotá - 2019. 1–201. Retrieved from http://rmcab.ambientebogota.gov.co/Pagesfiles/IA 200531 Informe Anual de Calidad del Aire Año 2019.pdf

SDP. (2020). Proyecciones de población. Retrieved from http://www.sdp.gov.co/sites/default/files/visor_proyecciones_sdp_v1.1_0.xlsm

Senanayake, I. P., Welivitiya, W. D. D. P., & Nadeeka, P. M. (2013). Remote sensing based analysis of urban heat islands with vegetation cover in Colombo city, Sri Lanka using Landsat-7 ETM+ data. Urban Climate, 5, 19–35. https://doi.org/10.1016/j.uclim.2013.07.004

Shen, L., Kyllo, J., & Guo, X. (2013). An Integrated Model Based on a Hierarchical Indices System for Monitoring and Evaluating Urban Sustainability. Sustainability, 5(2), 524–559. https://doi.org/10.3390/su5020524

Singh, P., Kikon, N., & Verma, P. (2017). Impact of land use change and urbanization on urban heat island in Lucknow city, Central India. A remote sensing based estimate. Sustainable Cities and Society, 32, 100–114. https://doi.org/10.1016/j.scs.2017.02.018

Soltani, A., & Sharifi, E. (2017). Daily variation of urban heat island effect and its correlations to urban greenery: A case study of Adelaide. Frontiers of Architectural Research, 6(4), 529–538. https://doi.org/10.1016/j.foar.2017.08.001

Ulpiani, G. (2021). On the linkage between urban heat island and urban pollution island: Three-decade literature review towards a conceptual framework. Science of the Total Environment, 751, 141727. https://doi.org/10.1016/j.scitotenv.2020.141727

UN-Habitat. (2019). Implementación de la Agenda 2030 y la Nueva Agenda Urbana. Retrieved from https://www.aciamericas.coop/xxiconferencia/wp-content/uploads/2019/12/05_Quintana-ONU-Habitat.pdf

UNDP. (2020). Goal 11: Sustainable cities and communities. Retrieved November 11, 2020, from https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-11-sustainable-cities-and-communities.html#targets

United Nations. (2007). Indicators of Sustainable Development: Guidelines and Methodologies (3th ed.). https://doi.org/10.1016/j.cirpj.2010.03.002

United Nations. (2019). World urbanization prospects The 2018 Revision. https://doi.org/10.18356/b9e995fe-en

Veeduría Distrital. (2018). Kennedy: Ficha Local. Retrieved from https://www.veeduriadistrital.gov.co/sites/default/files/files/Ficha Localidad Kennedy.pdf

Wang, S., Ma, Q., Ding, H., & Liang, H. (2018). Detection of urban expansion and land surface temperature change using multi-temporal landsat images. Resources, Conservation and Recycling, 128, 526–534. https://doi.org/10.1016/j.resconrec.2016.05.011

WHO. (2006). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/69477/WHO_SDE_PHE_OEH_06.02_eng.pdf;jsessionid=54263785E93420048269696C80477B40?sequence=1

Wu, X., Wang, G., Yao, R., Wang, L., Yu, D., & Gui, X. (2019). Investigating surface urban heat islands in South America based on MODIS data from 2003-2016. Remote Sensing, 11, 1212. https://doi.org/10.3390/rs11101212

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International Journal of Remote Sensing, 27(14), 3025–3033. https://doi.org/10.1080/01431160600589179

Yuan, F., & Bauer, M. E. (2007). Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. Remote Sensing of Environment, 106(3), 375–386. https://doi.org/10.1016/j.rse.2006.09.003

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. International Journal of Remote Sensing, 24(3), 583–594. https://doi.org/10.1080/01431160304987

Zhang, X., Estoque, R. C., & Murayama, Y. (2017). An urban heat island study in Nanchang City, China based on land surface temperature and social-ecological variables. Sustainable Cities and Society, 32(May), 557–568. https://doi.org/10.1016/j.scs.2017.05.005

Zhou, D., Xiao, J., Bonafoni, S., Berger, C., Deilami, K., Zhou, Y., … Sobrino, J. A. (2019). Satellite remote sensing of surface urban heat islands: Progress, challenges, and perspectives. Remote Sensing, 11(1), 1–36. https://doi.org/10.3390/rs11010048