UNIVERSIDAD POLITÉCNICA DE VALENCIA

DEPARTAMENTO DE ESTADÍSTICA, INVESTIGACIÓN OPERATIVA APLICADAS Y CALIDAD

MÁSTER EN INGENIERÍA DE ANÁLISIS DE DATOS, MEJORA DE PROCESOS Y TOMA DE DECISIONES

TRABAJO DE FIN DE MÁSTER

"ESTUDIO DE SIMILITUDES ENTRE PAÍSES Y ANÁLISIS DEL DESARROLLO MEDIANTE MODELOS BASADOS EN ESTRUCTURAS LATENTES CON DATOS FALTANTES"

Autor: Christian Enrique Tobar Estrella

Director: Manuel Zarzo Castelló

Valencia, 12 de Septiembre 2016

Agradecimientos

Me gustaría agradecer de todo corazón a mi tutor Manuel Zarzo, por su apoyo, enseñanza, compresión y paciencia en la exhaustiva labor de dirección y revisión de este trabajo.

Agradezco en gran medida a mis padres y hermanos que siempre me han impulsado a ser mejor persona en todo ámbito de mi vida.

Agradezco a mi tío Pablo E. por todo el apoyo y confianza depositados en mí.

Agradezco a mi novia María Fernanda C. por todo su amor, comprensión y ternura.

Dedico este trabajo a mi hijo Leandro Tobar B. que es el faro que guía mi vida.

RESUMEN

El presente Trabajo de Fin de Máster (TFM) plantea realizar un estudio de similitudes entre países y analizar el desarrollo mediante modelos basados en estructuras latentes con datos faltantes. Para ello se estructura en seis capítulos.

El primer capítulo introductorio aborda la motivación del trabajo, los objetivos y se describe la base de datos con la cual se ha llevado a cabo este TFM.

El capítulo 2 describe el marco conceptual que aborda el presente trabajo, contextualizado con la bibliografía pertinente, y se explican los modelos estadísticos que se van a realizar.

El capítulo 3 muestra los resultados de este trabajo, subdivididos en tres partes. La primera presenta un estudio exploratorio de países y variables mediante el modelo de análisis de componentes principales (PCA). Además se analiza la evolución temporal de las variables utilizando regresión PLS. En la segunda parte se analiza el índice de desarrollo humano a partir de modelos basados en estructuras latentes PCR (regresión en componentes principales) y PLS (regresión de mínimos cuadrados parciales). En el caso del modelo PLS se han considerado dos tipos de despliegue de la matriz tridireccional de datos, los cuales permiten un análisis complementario a la hora de evaluar el desarrollo humano. Finalmente, en una tercera parte se predice el Índice de Desarrollo Humano para los 35 países que no cuentan con ese dato. Además se ha analizado la capacidad predictiva de los 46 países que sí contaban con el valor de IDH pero que no se consideraron en los modelos de los capítulos anteriores por el alto porcentaje de datos faltantes. Para este conjunto de países, considerados a efectos de validación externa, se estimó el IDH por medio de dos métodos y se compararon los resultados mediante el estadístico RMSE.

En el capítulo 4 se presentan las conclusiones, y finalmente la bibliografía y anexos.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Motivación

En un contexto de globalización, es importante entender la dinámica mundial en la cual ciertos países tienden a conformar grupos o asociaciones en base a intereses políticos, económicos, ambientales, productivos, etc. Esto da motivo a interrogarse si acaso sólo los intereses definidos "en común" permiten estas asociaciones, y hasta qué punto los países tienen similitudes en ámbitos culturales, sociales o poblacionales que les permiten asociarse, parecerse y diferenciarse entre sí.

Este trabajo pretende analizar esa dinámica de similitudes y diferencias entre países, permitiendo una comparación a nivel mundial. Para ello se ha empleado el modelo de Análisis de Componentes Principales (PCA), el cual es una técnica exploratoria multivariante basada en estructuras latentes, que aporta una reducción de dimensionalidad, permitiendo visualizar gráficamente las relaciones entre países y variables.

Una pregunta de interés general es por qué existen países desarrollados y poco desarrollados. A pesar de que los países con bajo desarrollado realizan políticas públicas de intervención gubernamental y no gubernamental, las cuales no logran reducir la brecha, se plantea el interrogante del "porqué".

Con el objetivo de responder a esa pregunta, en este trabajo se analizará el grado de desarrollo a nivel mundial mediante modelos basados en estructuras latentes. Conocer mejor las similitudes y diferencias entre países resulta útil para discutir las políticas públicas. Es necesario que los gobiernos diseñen de forma óptima sus políticas públicas para conseguir los objetivos deseados y no se desvinculen de sus resultados finales. Por lo tanto, las metas de estas intervenciones deben estar enfocadas y dirigidas para alcanzar de forma clara y cuantificable una buena calidad de vida de la población.

Por último se analizará el Índice de Desarrollo Humano (IDH) a nivel mundial, para discutir las variables que mejor explican las diferencias en cuanto al desarrollo. Este índice no está disponible para muchos países. Por esta razón, se abordará en este trabajo la predicción del IDH para estos países mediante modelos de regresión basados en estructuras latentes: PCR y PLS.

1.2. Recopilación de datos

Se cuenta con una base de datos obtenida de internet, a partir de las páginas web de Naciones Unidas (data.un.org/), Banco Mundial (databank.bancomundial.org/data) y del Programa de las Naciones Unidas para el Desarrollo - PNUD (hdr.undp.org/en/data). La base de datos está compuesta por 120 indicadores de diferentes ámbitos (sociales, económicos, ambientales, poblacionales, de calidad de vida, igualdad, etc.), para todos los 234 países del mundo, en el período 2000 – 2013. No se han considerado los dos últimos años 2014 ni 2015 debido a la gran cantidad de datos faltantes.

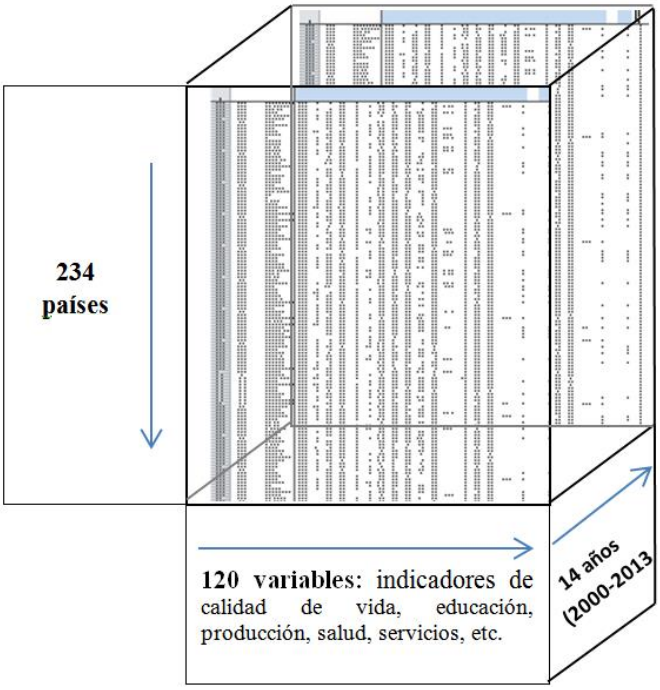

Figura 1. Estructura tridireccional de la base de datos

Entre las variables hay indicadores referidos a múltiples ámbitos: acceso y calidad de servicios básicos (agua, sanitarios, educación), salud (VIH, tuberculosis, desnutrición, cuidado durante el embarazo, atención en parto), pobreza (población que vive con menos de un dólar por día), ambientales (emisión de dióxido carbónico), ingresos (PIB constante, ingreso per cápita), ubicación geográfica, etc.

De estos 393.120 posibles registros (en caso de tener la base de datos completa: 234 x 120 x 14), se cuenta con 90.643 registros, lo cual representa un 23% de la matriz completa. Esta enorme cantidad de datos faltantes puede deberse a una o varias causas, las cuales han sido modelizadas por Rubin (1976) en función de tres posibilidades:

i) Modelo MCAR. La falta de información es debida a una pérdida completamente aleatoria, con igual probabilidad de pérdida para todos los datos.

ii) Modelo MAR. La pérdida de información es aleatoria, cuya probabilidad está en relación a la información observada en las otras variables. Por ejemplo, un país con valores bajos de gasto público en áreas sociales, previsiblemente tendrá un bajo ingreso público total. Pero si realmente ese país tiene un ingreso público alto, este último dato podría no ser publicado intencionadamente por sus organismos gubernamentales para no poner de manifiesto una mala gestión.

iii) Modelo MNAR. La pérdida de información no es aleatoria, depende de los valores de la información "pérdida". Ciertos valores pueden resultar incómodos o poner en evidencia una mala gestión gubernamental, por lo que es posible que tiendan a ocultarse (es decir, no se hacen públicos). Por ejemplo, un país con tasas altas de analfabetismo es posible que no haga pública esta información. Así pues, la presencia de datos faltantes puede hacer sospechar que pueden ser atípicos.

1.3. Objetivos

El objetivo general de este trabajo es entender mejor la relación entre indicadores (de calidad de vida, educación, producción, salud, servicios, etc.) a nivel mundial, para discutir los factores que se relacionan con el desarrollo humano de un país. También se pretende obtener modelos predictivos del índice de desarrollo humano (IDH) para aquellos países en los cuales no está disponible, y para discutir las variables que mejor explican las diferencias en cuanto al desarrollo. Con todo esto se pretende aportar criterios que favorezcan el desarrollo de políticas públicas útiles para potenciar el desarrollo humano de los países, en especial de los menos desarrollados.

Los objetivos específicos que se plantean son los siguientes:

i. Realizar un análisis exploratorio de las variables para estudiar las estructuras de correlación que definen las similitudes o diferencias entre los países.
ii. Aplicar técnicas de regresión para predecir el IDH en función de los indicadores, empleando métodos estadísticos basados en estructuras latentes (PCR y PLS). Discusión de los resultados obtenidos con los distintos modelos.
iii. Predecir el IDH en aquellos países que no cuentan con ese índice.

1.4. Países e indicadores descartados por falta de información

La matriz tridireccional inicial (Figura 1) está caracterizada por un elevado número de datos faltantes. Por ello, en primer lugar es necesario discutir cómo actuar ante esta situación. Se propone descartar aquellas observaciones (países) y variables con alta cantidad de valores faltantes. Se ha decidido eliminar aquellos países que acumuladamente en los 14 años no cuenten con al menos un 50% de datos en las 120 variables (eliminación de filas). Con este criterio, el número inicial de 234 países se reduce a 139.

Posteriormente, se decide eliminar las variables que no cuentan con al menos el 30% de datos. De estas variables eliminadas, al inspeccionarlas se observa que en su mayoría eran deficientes en datos debido a alguno de los siguientes tres aspectos: i) a nivel de un sector entero (educación, salud, gasto público, etc.), ii) por región geográfica (como “África Subsahariana”, “Medio Oriente y norte de África”, entre otros), y iii) con información temporal parcial (por ejemplo, datos disponibles de un solo año). De estas variables eliminadas, el 37% tenían más del 85% de datos faltantes. Con esta reducción se pasa de 120 a 48 variables.

Ambos criterios considerados (50% al menos de datos en los países y 30% al menos en las variables) pueden parecer un poco arbitrarios, pero han sido establecidos tras un estudio preliminar. Se probaron varias posibles combinaciones entre eliminación de filas y columnas, según el cual la combinación elegida (50% y 30%) mostró que permitía mantener la mayor cantidad de información posible para el mayor número de países. El criterio empleado también permitía contar con el 60% de países a nivel mundial, lo cual es un valor razonablemente alto. Las Naciones Unidas ofrecen el valor del índice de Desarrollo Humano para un número similar de países, por lo cual el criterio empleado parece razonable.

Con el procedimiento indicado, la base de datos inicial se reduce a una matriz tridireccional con 139 países, 48 variables y 14 años, como se puede observar en la siguiente figura.

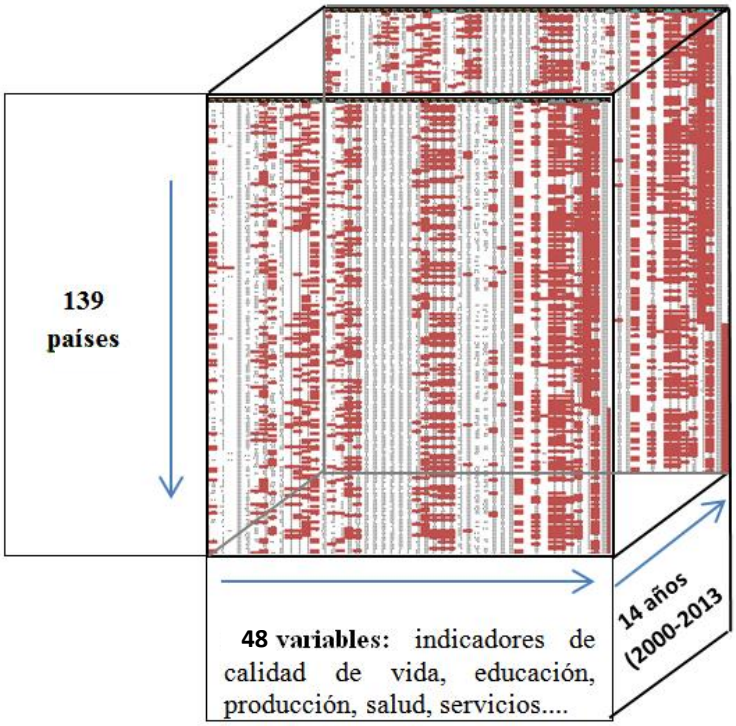

Figura 2. Matriz tridireccional tras eliminar los países y variables con pocos datos

La lista de las 48 variables se presenta en la Tabla 12 del anexo I. Esta matriz todavía contiene un 19,6% de datos faltantes. No obstante, los programas estadísticos ProSensus y SIMCA-P empleados en el presente trabajo son capaces de manejar matrices con valores faltantes, por lo que no ha sido necesario considerar métodos alternativos para la imputación de los valores desconocidos. Ambos programas, que permiten realizar análisis de componentes principales y regresión PLS, utilizan el algoritmo iterativo NIPALS (NonLinear Iterative Partial Least Squares), el cual se ha diseñado convenientemente para abordar el caso de matrices con cierta cantidad de datos faltantes, como se explica en la sección 2.2.4.

2. MARCO TEÓRICO

2.1. Conceptualización del problema

2.1.1. Contexto mundial

A lo largo de la historia, los países se han clasificado tanto por sus ideologías político-económicas como por su grado de desarrollo. En el último siglo, las principales clasificaciones son las siguientes:

i) Bloques económico-políticos de la guerra fría

Una clasificación propuesta por el economista Alfred Sauvy (1952) catalogaba los países según su "alineación" en bases a los siguientes bloques: "Primer Mundo" conformado por Estados Unidos, Francia, Reino Unido y sus aliados; "Segundo Mundo" (Unión Soviética, China y sus aliados); y "Tercer Mundo" (Latinoamérica, India, África, Sur de Asia y Medio Oriente).

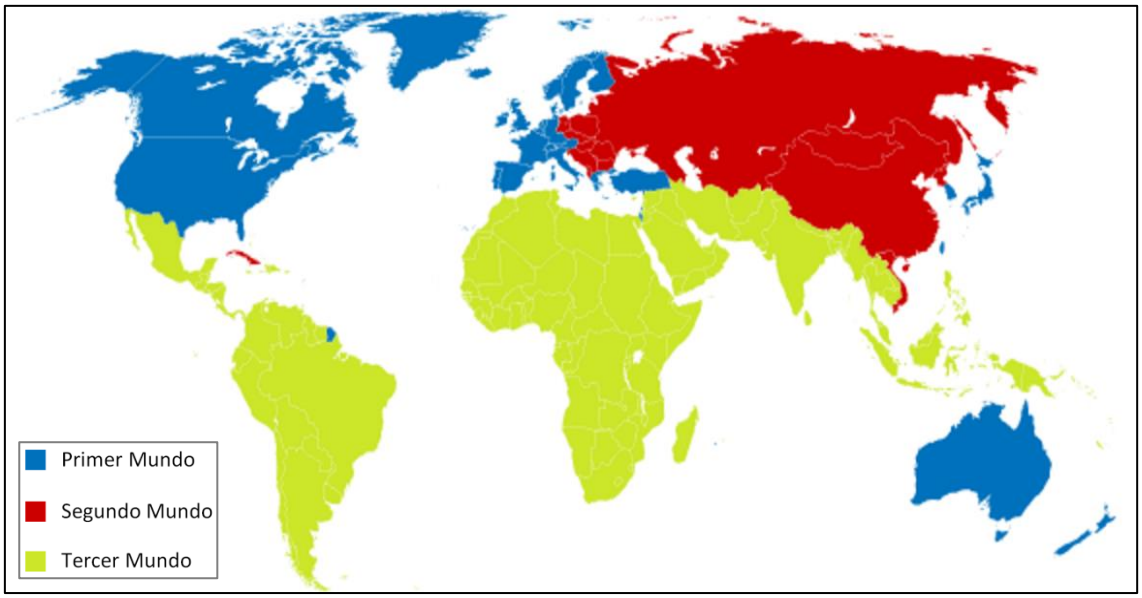

Figura 3. Clasificación de países por bloques económico-políticos en la guerra fría

ii) Desarrollo económico

También se han propuesto términos como "países en vías de desarrollo" por parte del Banco Mundial, y "economías emergentes", propuestos por el Fondo Monetario Internacional, los cuales se refieren a países con economías menos desarrolladas que Estados Unidos, Gran Bretaña y algunos países de Europa. Se denominan emergentes ya que presentan un rápido crecimiento económico y una progresiva industrialización, y cuentan con riqueza de materias primas, de recursos, de una relativa estabilidad política y de una elevada demanda interna.

Globalización[1]

La actual dinámica mundial se encuentra inmersa en la conocida "globalización". En el artículo "replanteamiento de la globalización", publicado por las Naciones Unidas en un contexto de los objetivos de desarrollo del segundo milenio, se define la globalización en los siguientes términos (fuente: http://www.un.org/es/aboutun/booklet/globalization.shtml).

La globalización es un fenómeno inevitable en la historia humana que ha acercado el mundo a través del intercambio de bienes y productos, información, conocimientos y cultura. En las últimas décadas, la integración mundial ha cobrado velocidad de forma espectacular debido a los avances sin precedentes en la tecnología, las comunicaciones, la ciencia, el transporte y la industria.

Si bien la globalización es a la vez un catalizador y una consecuencia del progreso humano, es también un proceso caótico que requiere ajustes y plantea desafíos y problemas importantes.

Entre las críticas de los efectos de la globalización se encuentra la integración económica, la cual se produce cuando los países reducen los obstáculos, como los aranceles de importación, y abren su economía a la inversión y al comercio con el resto del mundo. Los detractores se quejan de que las disparidades que se producen en el sistema comercial mundial de hoy perjudican a los países en desarrollo.

Los defensores de la globalización señalan que países como China, Vietnam, India y Uganda que se han abierto a la economía mundial han reducido notablemente la pobreza. Los críticos sostienen que el proceso ha significado la explotación de gente en los países en desarrollo, produciendo perturbaciones masivas y aportando pocos beneficios.

Para que todos los países puedan beneficiarse de la globalización, la comunidad internacional debe seguir esforzándose por reducir las distorsiones en el comercio internacional que favorecen a los países desarrollados y por crear un sistema más justo.

Muchos países de África no se han beneficiado de la globalización. Sus exportaciones han seguido limitándose a unos cuantos productos básicos. Algunos expertos señalan que las deficiencias de las políticas y la infraestructura, la debilidad de las instituciones y la corrupción en los organismos públicos han marginado a diversos países. Otros creen que algunos aspectos geográficos y climáticos desfavorables han dejado a algunos países fuera del crecimiento mundial.

Entre los países que se han beneficiado de la globalización está India, que ha reducido a la mitad la tasa de pobreza en las últimas dos décadas. En China, la reforma ha propiciado la mayor disminución de la pobreza de la historia. El número de pobres en las zonas rurales pasó de 250 millones en 1978 a 34 millones en 1999.

[1] Replanteamiento de la globalización, en un contexto de los objetivos de desarrollo del milenio, publicaciones Web de Naciones Unidas.

Ingreso per cápita y población

Un indicador frecuentemente empleado es la relación entre el PIB (Producto Interno Bruto) de un país dividido por su población (PIB per cápita). Si bien este indicador ha sido utilizado en las últimas décadas como una medida de crecimiento y desarrollo económico de un país, este indicador no mide el nivel de "bienestar" real de la población, ya que existen desigualdades en la distribución de la riqueza de un país. Es por esto que otros indicadores como el Índice de Desarrollo Humano han sido propuestos para cuantificar en mejor medida el desarrollo no solo económico, sino del nivel de vida de la población de cada país.

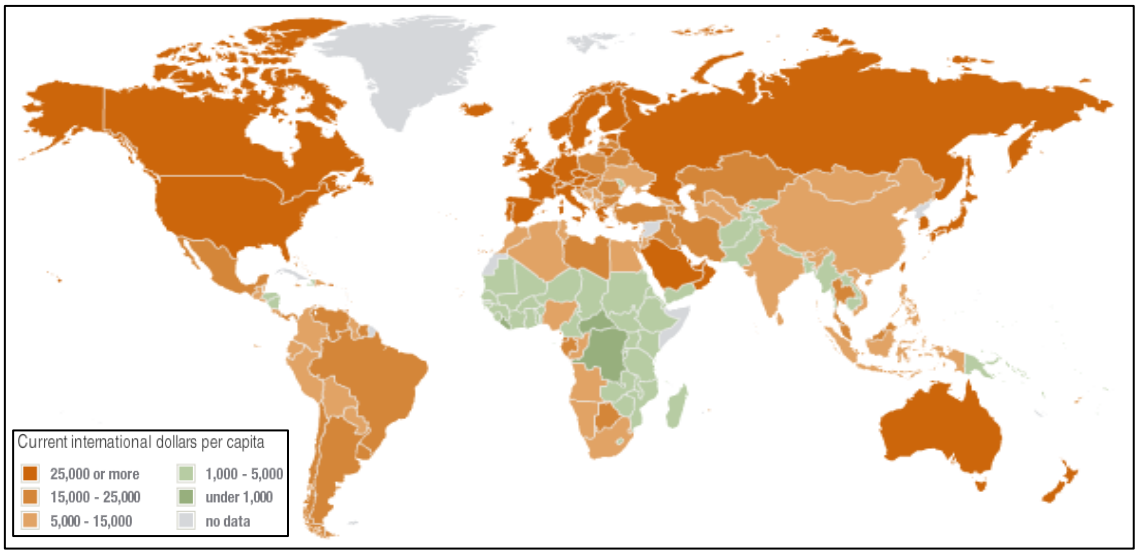

Figura 4. PIB per cápita (corregido con la Paridad de Poder de Compra) por país a nivel mundial, datos de 2013.
Fuente: Mapa de datos, Aplicación web del Fondo Monetario Internacional (FMI)

Un punto importante en el contexto mundial es la consideración del crecimiento poblacional. En 2013 se observa a Estados Unidos, Brasil, India y China como países con gran número de habitantes, tal como se ilustra en la siguiente figura:

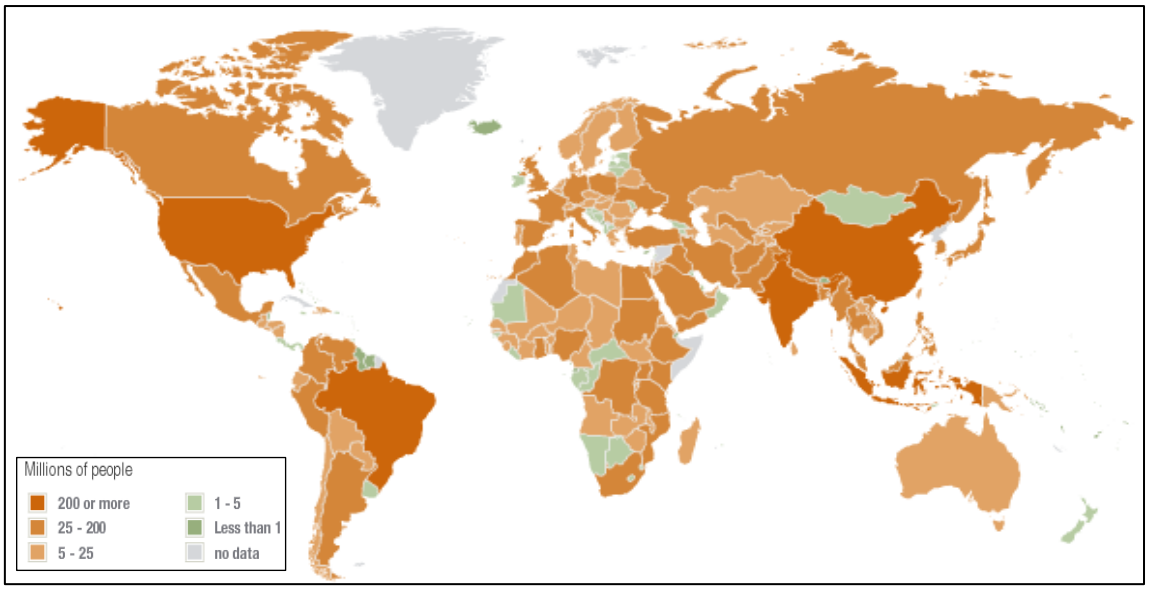

Figura 5. Población por país a nivel mundial, datos de 2013
Fuente: Mapa de datos, Aplicación web del Fondo Monetario Internacional (FMI)

2.1.2. Desarrollo

Si bien el desarrollo se entiende a grandes rasgos como una “mejora”, este concepto ambiguo puede tener distintos enfoques, dando lugar a muchas interpretaciones de “desarrollo”. Por ejemplo, se podría entender como desarrollo a la mejora de un país en uno o varios temas a nivel global o local. También se podría entender el desarrollo de forma individual como la mejora de una persona en sus hábitos, psicología, forma de ser o actuar.

En este trabajo se considera el desarrollo como la mejora de la sociedad en su conjunto. Esta consideración recae en la línea conceptual de las Naciones Unidas, que ha declarado desde su primer Informe sobre Desarrollo Humano realizado en 1990 que “la verdadera riqueza de una nación está en su gente”.

Tanto a nivel nacional como local, el desarrollo es el objetivo de muchas sociedades. En su búsqueda, gobiernos y organismos internacionales intervienen en las actividades que se generan en un territorio. Estas intervenciones se realizan a través de políticas públicas o asistencia internacional, las cuales sirven como herramienta con la que cuentan los gobiernos, tanto a escala nacional como local, para implementar soluciones a problemáticas específicas.

En cuanto a una definición específica, autores como Boisier (2001) muestran que en torno al desarrollo se ha construido una multiplicidad de significados que reclaman identidad única en relación al adjetivo “desarrollo”. De hecho, se han propuesto conceptos como desarrollo territorial, desarrollo regional, desarrollo local, desarrollo endógeno, desarrollo sostenible, etc.

Para entender el desarrollo se debe analizar el “porqué” del surgimiento de esta teoría de pensamiento. Por ejemplo, en base a la experiencia del desarrollo en Latinoamérica, Joseph Stiglitz (Premio Nobel de Economía, ex Vicepresidente del Banco Mundial) sugería en la década de los 90 que se debía reexaminar y ampliar los conocimientos sobre la economía del desarrollo (Stiglitz, 1998).

Para este trabajo conviene realizar una síntesis de los principales conceptos sobre el desarrollo y la interrelación entre ellos, con el fin de entender el desarrollo como la suma de los conceptos inmersos en su definición.

Desarrollo endógeno

Para Bosier (2001), el desarrollo endógeno nace como reacción al pensamiento y a la práctica dominante en materia de desarrollo territorial en las décadas de 1950 y 1960, enmarcados en el paradigma industrial fordista y en la difusión de las innovaciones y de los impulsos de cambio.

Para Vázquez-Barquero (2004), la teoría del desarrollo endógeno comienza a ser construida a partir de los años 1980 en base a experiencias de esa época especialmente en Europa, USA y el mundo desarrollado.

Para Boisier (1993), la endogeneidad del desarrollo regional habría que entenderla como un fenómeno que se presenta por lo menos en cuatro planos que se entrecruzan entre sí:

i) Plano político, en el cual el desarrollo implica una creciente capacidad regional para tomar las decisiones relevantes.

ii) Plano económico, que se refiere a la apropiación y reinversión regional de parte del excedente a fin de diversificar la economía regional.

iii) Plano científico y tecnológico, es decir, la capacidad interna de un territorio organizado para generar sus propios impulsos tecnológicos de cambio, capaces de provocar modificaciones cualitativas en el territorio.

iv) Plano de la cultura, como una matriz generadora de la identidad socioterritorial.

Por su parte, Garofoli (1995), uno de los más notables exponentes del "nuevo regionalismo" europeo, define el desarrollo endógeno como *"la capacidad para transformar el sistema socio -económico; la habilidad para reaccionar a los desafíos externos; la promoción de aprendizaje social; y la habilidad para introducir formas específicas de regulación social a nivel local que favorecen el desarrollo de las características anteriores".*

Romer (1994) destaca que el crecimiento económico es un resultado endógeno de un sistema económico y no del resultado de fuerzas que inciden desde fuera. Existe una creencia neoclásica de que los países a largo plazo convergen en el desarrollo.

Stiglitz (2002) considera que "el desarrollo abarca no sólo recursos y capital sino una transformación de la sociedad". El propio Presidente del Banco Mundial, James Wolfensohn, ha sostenido que "la auto-conciencia y el orgullo que viene de la identidad cultural es una parte esencial del empoderamiento de las comunidades para tomar en sus manos su propio destino" (Banco Mundial, 2004).

2.1.3. Políticas públicas y desarrollo

Para Mesa (2014) la implementación de políticas públicas encaminadas a lograr objetivos sociales y económicos en una región particular, puede contribuir al desarrollo de una sociedad si se tiene en cuenta como metas la generación de empleo, la atracción de turistas como fuente de ingresos, la mejora de infraestructuras y de viviendas, etc.

Si bien la política pública es un pilar para el desarrollo, su conceptualización se encuentra todavía en discusión. Entre las principales conceptualizaciones se encuentran las siguientes:

Tabla 1. Definición de política pública y desarrollo. Elaboración: Mesa (2014)

Autor	Definición de Política Pública
Subirats (1989)	Son las actividades de las instituciones de gobierno, actuando directamente o a través de agentes, y que van dirigidas a tener una influencia determinada sobre la vida de los ciudadanos.
Repetto (2001)	Es el resultado de la interacción entre actores sociales y estatales, modelados por marcos institucionales.
Piñango (2003)	Son proposiciones gubernamentales sobre la mejor forma de lograr determinados objetivos sociales.
Dye (2008)	Es todo lo que los gobiernos deciden hacer o no hacer.
Maggiolo (2007)	Son las disposiciones del Estado para atender determinadas realidades que afectan directa e indirectamente a la sociedad, sean del tipo social, política o económica.
De Kostka (2013)	Es el programa de acción de una autoridad gubernamental, que afecta a un sector de la sociedad o bien a un espacio geográfico determinado.
Borrás (2014)	Los gobiernos a menudo implementan programas con el objetivo de cambiar resultados económicos o sociales de los individuos.

Del resumen realizado por Mesa (2014) se puede deducir que la mayoría de autores concuerdan en que las políticas públicas ayudan al desarrollo tanto nacional como local, y que la acción público-privada potencia de mejor manera el desarrollo.

Elizalde (2003a) y Mesa (2014) señalan que desde los años 80 existe un cambio en el modo de planificar el desarrollo, ya que la planificación regional o local solía ser, antes de las últimas décadas, una réplica de la nacional.

En cuanto al modelo de planificación, varios autores coinciden en que se debe realizar en dos niveles, tanto en lo nacional como en lo local. A nivel nacional con planes estratégicos que deben estar ligados al surgimiento de instituciones de financiación flexibles, la integración entre agentes públicos y privados y la innovación en las formas de gestión y organización productiva. Y a nivel local, se debe tener una visión más estratégica del problema del desarrollo, cuya actividad debe reorientarse para incidir en la colaboración con los agentes económicos y financieros con el objetivo de encontrar diferencias competitivas vinculadas al territorio y a la utilización de recursos endógenos, auspiciando la concertación estratégica entre lo público y lo privado (Elizalde, 2003b).

Retos en políticas públicas para promover el desarrollo

Si bien el desarrollo siempre es deseado, históricamente han primado en las políticas públicas las orientaciones y criterios que obedecen a una planificación centralista tecnocrática y cortoplacista, que busca una alta rentabilidad en relación a su costo, y resultados que permitan un beneficio político significativo.

En esta línea, Morales (2013) muestra como ejemplo la política pública en México que, si bien pretende satisfacer las necesidades de la población, es un sistema deficiente y costoso. De hecho, en un análisis realizado en el período 2001 - 2012, del 100% de los recursos destinados a gasto público, únicamente el 11% fueron directos a políticas públicas de inversión, mientras que el 89% restante era de gasto corriente. Estas cifras muestran la falta de planeamiento en

la financiación para políticas públicas, que incluye precisamente cuestiones sobre desarrollo local y autosostenibilidad.

En América Latina, al igual que muchos países en vías del desarrollo, existe una alta preocupación por la posible corrupción en la implantación de políticas públicas, las cuales benefician en gran medida a intereses personales, familiares o ciertas élites. Esto afecta sobremanera al logro del desarrollo y aumenta las brechas ya existentes en cuanto a desigualdad y pobreza. Para combatir esta corrupción resulta conveniente que el desarrollo de políticas públicas se realice con la participación ciudadana, lo que ayuda a tener un empoderamiento de las acciones a favor del desarrollo territorial.

Según Murga (2006), es necesario que los actores externos apoyen la permanencia del desarrollo local endógeno (ONG, gobiernos centrales u otros organismos) a efectos de: i) elegir los proyectos y territorios para enfocar esfuerzos y recursos, ii) escuchar empáticamente los problemas, iii) estimular, ayudar y motivar a los líderes y actores locales, iv) comunicación mediante conceptos y palabras sencillas, v) ayudar a mediar en los inevitables conflictos locales, vi) entrega de apoyo técnico experto a escala de los problemas a resolver, vii) comprender y respetar las estructuras de poder locales, viii) ayudar a difundir el conocimiento local generado en el proceso de desarrollo, y ix) aceptar que las comunidades se apropien de sus logros colectivos.

2.1.4. Desarrollo humano[2]

La página web del Programa de las Naciones Unidas para el Desarrollo (PNUD) recopila diversos textos donde se discute el concepto de Desarrollo Humano. Según Amartya Sen, Premio Nobel de Economía en 1998, el desarrollo humano forma parte de la esencia misma del desarrollo: el aumento de la riqueza de la vida humana en lugar de la riqueza de la economía en la cual viven los seres humanos.

Esa fue la visión inicial y sigue siendo el punto de vista de los autores del primer Informe sobre Desarrollo Humano de las Naciones Unidas, realizado por Mahbub ul-Haq de Pakistán y Amartya Sen de la India, junto con otros importantes ideólogos en este campo. Su concepción ha orientado la redacción del Informe del Desarrollo Humano durante 20 años, así como más de 600 informes nacionales sobre desarrollo humano —elaborados a partir de investigaciones locales y publicados por sus respectivos países— y también múltiples informes de carácter regional apoyados por las oficinas regionales del PNUD.

El concepto de desarrollo humano se centra en los fines y no en los medios de progreso. El objetivo verdadero del desarrollo debería apuntar a crear un ambiente propicio para que la gente disfrute de una vida larga, saludable y creativa. Esto parece una verdad sencilla, pero muchas veces se pasa por alto por dar prioridad a cuestiones más inmediatas.

El desarrollo humano denota tanto el proceso de ampliar las opciones de las personas como la optimización de su bienestar. Los aspectos cruciales del desarrollo humano son: una vida

[2] El concepto de Desarrollo Humano es la recopilación de lo publicado en la página Web del Programa de las Naciones Unidas para el Desarrollo.

prolongada y saludable, la educación y un nivel de vida digno. Otras cuestiones incluyen las libertades sociales y políticas.

El concepto distingue dos partes. Por un lado está la formación de las capacidades humanas, tal como la mejora en la salud o la educación. La otra parte comprende disfrutar las capacidades adquiridas, ya sea para trabajar o para disfrutar del tiempo libre.

En reiteradas ocasiones, el desarrollo humano se ha malinterpretado y se ha confundido con los siguientes enfoques:

- El crecimiento económico es un medio y no un fin del desarrollo. Además, un PIB alto no significa necesariamente que haya progreso en términos de desarrollo humano. La experiencia mundial ha demostrado que los ingresos y el desarrollo humano no van siempre juntos. Algunos países tienen niveles relativamente altos de desarrollo humano en comparación con sus ingresos y viceversa.
- Las teorías de formación en capital humano y desarrollo de recursos humanos ven a las personas como un medio de obtener mayores ingresos y más riqueza en lugar de verlas como un fin. Estas teorías consideran a los seres humanos como factores para aumentar la producción.
- El enfoque del bienestar humano considera a las personas como beneficiarios, y no como participantes del proceso de desarrollo.
- El enfoque de las necesidades básicas se concentra en el conjunto de bienes y servicios fundamentales que necesita la población: alimento, vivienda, indumentaria, salud y agua. Se dedica al suministro de estos bienes y servicios y no a lo que implican para las personas.

Por lo tanto, el concepto de desarrollo humano es holístico y sitúa a las personas en el centro de todos los aspectos del proceso de desarrollo. Además, el desarrollo humano se trata de un concepto en constante evolución, cuyas herramientas analíticas se adaptan a los cambios que ocurren en el mundo. El paradigma del desarrollo humano es aplicable a todos los países, ricos y pobres, y a todos los seres humanos.

2.1.5. Cuantificación del desarrollo

Para medir el grado de desarrollo de un país, en este trabajo se considera el Índice de Desarrollo Humano (IDH), propuesto por el Programa de las Naciones Unidas para el Desarrollo, el cual permite comparar a los diferentes países en función de sus indicadores tanto sociales como económicos.

El IDH cuantifica de forma sintética los logros medios obtenidos en las dimensiones fundamentales del desarrollo humano: (i) tener una vida larga y saludable, (ii) adquirir conocimientos, y (iii) disfrutar de un nivel de vida digno. Según la definición del IDH publicada en la web del PNUD (http://hdr.undp.org/en/content/human-development-index-hdi), este indicador se calcula como promedio de los índices normalizados de cada una de las tres dimensiones. El IDH fluctúa entre 0 y 1, siendo 0 un valor correspondiente a un país sin desarrollado alguno y el valor 1 indica un país con un desarrollo muy alto.

La dimensión de la salud se evalúa según la esperanza de vida al nacer. La de educación se mide por los años en promedio de escolaridad de los adultos con 25 años o más, y por los años esperados de escolaridad de los niños en edad escolar. La dimensión del nivel de vida se mide conforme al PIB per cápita. Las puntuaciones de los tres índices dimensionales del IDH se agregan posteriormente a un índice compuesto utilizando la media.

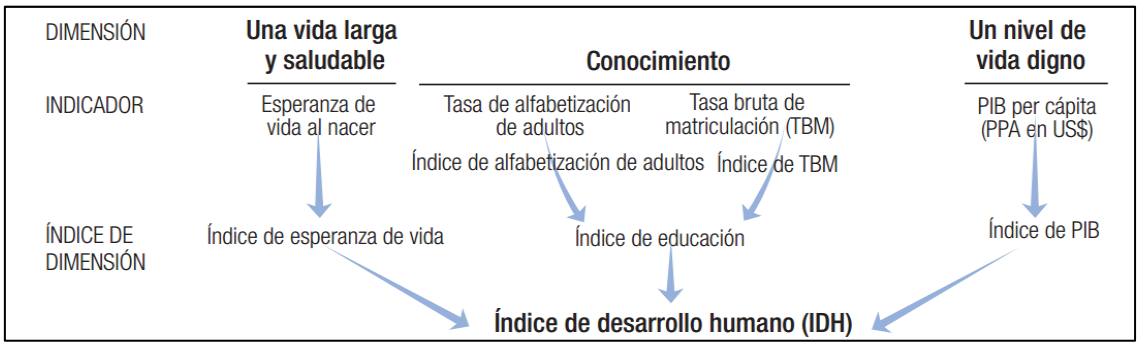

Figura 6. Esquema del cálculo del Índice de Desarrollo Humano

Cálculo del índice[3]

Tal como se ha explicado, para calcular el IDH se crea primero un índice para cada una de las 3 dimensiones. Para calcular estos índices (esperanza de vida, educación y PIB), se eligen los valores mínimos y máximos (límites) para cada uno de los indicadores básicos.

El desempeño en cada dimensión se expresa como un valor entre 0 y 1 tras aplicar la siguiente fórmula general:

$$\text{Índice de dimensión} = \frac{\text{valor real} - \text{valor mínimo}}{\text{valor máximo} - \text{valor mínimo}}$$

Los límites para calcular el IDH se indican en la siguiente tabla:

Tabla 2. Límites de los indicadores del cálculo del IDH

Indicador	Valor máximo	Valor mínimo
Esperanza de vida al nacer (en años)	85	25
Tasa de alfabetización de adultos (%)	100	0
Tasa de matriculación bruta combinada (%)	100	0
PIB per cápita (USD)	40.000	100

a. Cálculo del índice de esperanza de vida

El índice de esperanza de vida mide el logro relativo de un país en la esperanza de vida al nacer. En el caso de Brasil, por ejemplo, con una esperanza de vida de 70,8 años en 2004, el índice de esperanza de vida es 0,764.

$$\text{Índice de esperanza de vida} = \frac{70{,}8 - 25}{85 - 25} = 0{,}764$$

[3] Informe de Desarrollo Humano 2006. Programa de las Naciones Unidas para el Desarrollo (PNUD). Pág. 393-394.

b. Cálculo del índice de educación

El índice de educación mide el logro relativo de un país en la alfabetización de adultos y la matriculación bruta combinada en escuelas primarias, secundarias y terciarias. En primer lugar se calcula el índice de alfabetización de adultos y el índice de matriculación bruta combinada. A continuación, estos dos índices se combinan para crear el índice de educación, con una ponderación de dos tercios para la alfabetización de adultos y de un tercio para la matriculación bruta combinada.

En el caso de Brasil, con una tasa de alfabetización de adultos del 88,6% en 2004 y una tasa de matriculación bruta combinada del 86% en 2004, el índice de educación es de 0,876.

$$\text{Índice de alfabetización de adultos (IAA)} = \frac{88{,}6 - 0}{100 - 0} = 0{,}886$$

$$\text{Índice de matriculación bruta (IMB)} = \frac{86 - 0}{100 - 0} = 0{,}857$$

$$\text{Índice de educación} = \frac{2}{3}\ (\text{IAA}) + \frac{1}{3}(\text{IMB}) = \frac{2}{3}\ (0{,}886) + \frac{1}{3}(0{,}857) = 0{,}876$$

c. Cálculo del índice de PIB

Para calcular el índice de PIB se utiliza el valor ajustado del PIB per cápita (en US$), parámetro denominado PPA (PIB per cápita por paridad del poder adquisitivo). En el IDH, los ingresos se ajustan porque para lograr un nivel respetable de desarrollo humano no son necesarios ingresos ilimitados. En consecuencia, se utiliza el logaritmo de los ingresos.

En el caso de Brasil, siendo el PIB ajustado per cápita de 8.195 $ (PPA en US$) en 2004, el índice de PIB es de 0,735.

$$\text{Índice del PIB} = \frac{\log(8.195) - \log(100)}{\log(40.000) - \log(100)} = 0{,}735$$

d. Cálculo del IDH

Una vez calculados los índices de dimensión, se calcula el promedio simple de los tres índices de dimensión.

IDH = 1/3 (índice de esperanza de vida) + 1/3 (índice de educación) + 1/3 (índice de PIB)
= 1/3 (0,764) + 1/3 (0,876) + 1/3 (0,735) = 0,792

El IDH se creó para hacer hincapié en que las personas y sus capacidades —y no el crecimiento económico por sí mismo— deben ser el criterio más importante para evaluar el desarrollo de un país. El IDH también puede usarse para cuestionar las decisiones normativas nacionales, comparando cómo dos países con el mismo nivel de PIB per cápita obtienen resultados diferentes en materia de desarrollo humano. Estos contrastes pueden impulsar el debate sobre las prioridades normativas de los gobiernos.

El IDH simplifica y refleja sólo una parte de lo que entraña el desarrollo humano, ya que no contempla las desigualdades, la pobreza, la seguridad humana ni el empoderamiento. Los valores del IDH para el 2014 se reflejan en la siguiente figura:

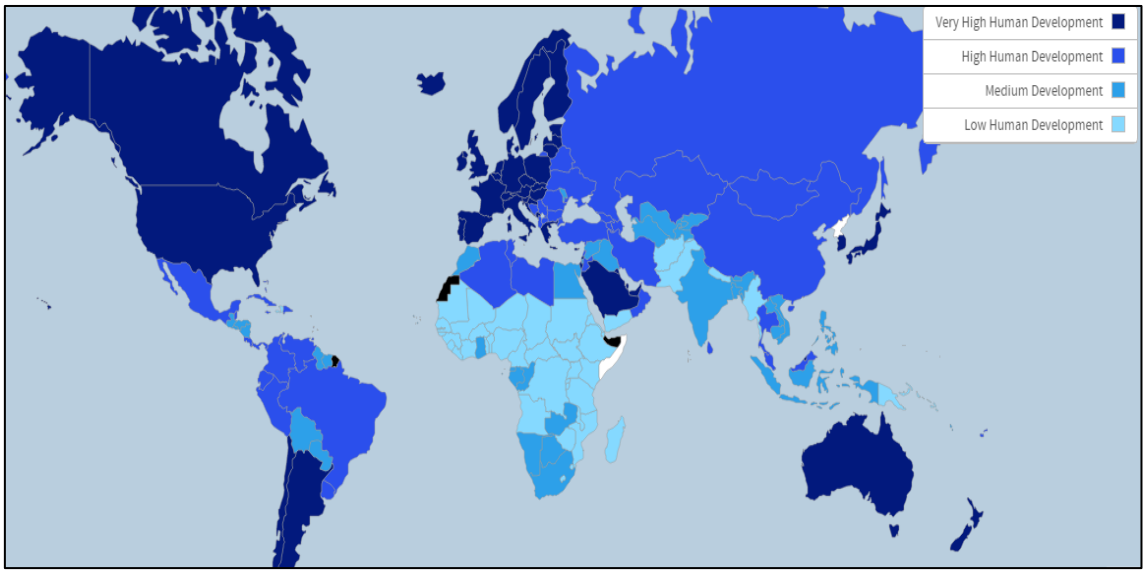

Figura 7. Clasificación de los países en función del Índice de Desarrollo Humano para 2014

2.2. Modelos estadísticos basados en estructuras latentes

Describir la situación de un país requiere tener en cuenta simultáneamente diversos parámetros. Las variables que describen un país en relación a la calidad de vida, población, igualdad de género, desarrollo, producción, etc. no se deben analizar de forma univariante, ya que además de infrautilizar información valiosa, se tendría que realizar un análisis gráfico de la relación entre variables con múltiples gráficos de dos dimensiones. Por ello, conviene emplear métodos exploratorios de datos basados en análisis multivariante que utilizan todos los datos disponibles para extraer convenientemente la información que contienen.

Según Peña (2002), el análisis de datos multivariantes tiene por objeto el estudio estadístico de varias variables medidas en los individuos de una población. La utilización de estas técnicas estadísticas permite:

1. Resumir el conjunto de variables en unas pocas nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información.
2. Encontrar grupos en las observaciones (clústers), si existen.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables.
5. Identificar datos atípicos.

El coeficiente de correlación, el cual se calcula a partir de la covarianza, se emplea habitualmente para cuantificar la relación lineal entre dos variables. El primer paso para caracterizar datos multivariantes es describir cada variable (media, varianza) y, posteriormente, comprender la estructura de dependencia que existe entre ellas mediante la matriz de varianzas - covarianzas. Esta matriz resume la información multivariante contenida en los datos.

El problema de condensar la información de un conjunto de variables se resuelve construyendo nuevas variables denominadas latentes, las cuales sintetizan la información contenida en las variables originales. Existen distintos métodos exploratorios para conseguir este objetivo. Con variables continuas, como es el caso de este trabajo, el método más utilizado es el Análisis de Componentes Principales (PCA), el cual fue propuesto por primera vez por Pearson (1901) y Hotelling (1933).

Las componentes principales permiten representar adecuadamente los datos. A partir de ellas se pueden realizar gráficos en pocas dimensiones con la mínima pérdida de información, para entender la estructura subyacente de los datos. En el presente trabajo se ha empleado este método, ya que se desea aprovechar las ventajas del PCA para poner de manifiesto las semejanzas entre países y entre variables.

2.2.1. Modelo de Análisis de Componentes Principales - PCA

El PCA es una técnica estadística propuesta originariamente por Pearson (1901), la cual busca líneas y planos que ajustan lo mejor posible a un sistema de puntos en el espacio. Por ello, la técnica sintetiza información y reduce el número de dimensiones para facilitar la interpretación de la relación entre variables y observaciones.

Es una técnica muy útil cuando existen altas correlaciones entre variables, ya que se pueden ajustar las observaciones con pocas direcciones o líneas por el criterio de mínimos cuadrados. Estas direcciones de variabilidad, denominadas "*componentes principales"*, explican gran parte de la variabilidad de los datos analizados (Jolliffe, 2002).

El cálculo de las componentes principales se realiza de tal forma que la primera recoge la mayor proporción posible de la variabilidad original de los datos; la segunda componente explica la máxima variabilidad posible no recogida por la primera, y así sucesivamente.

Una ventaja del PCA es que las componentes principales son combinaciones lineales de las variables originales. Todas las componentes principales son independientes entre sí, es decir, son ortogonales (Esbensen y Geladi, 1987).

Una vez calculadas las direcciones de máxima varianza (componentes principales), pueden proyectarse las observaciones sobre estas direcciones, las cuales cuentan con un valor de la proyección de la observación sobre una componente que se denomina "*score"* y una dirección sobre la que se proyecta la observación que es definida por unos pesos (*loadings*). Así pues, los resultados de un PCA se discuten en términos de puntuaciones o "scores" y a partir de los pesos o "loadings".

Las proyecciones de las observaciones sobre cada una de las componentes da como resultado distintos "vectores de scores" denominados variables latentes, las cuales pueden considerarse como nuevas variables que contienen información subyacente de los datos.

Procedimientos de validación del PCA

Para interpretar los resultados del modelo PCA hay que determinar primero la relevancia de las componentes del modelo e identificar anomalías en las observaciones, es decir, patrones atípicos en algún país.

Se trata de encontrar el número de componentes que modelicen la variabilidad de los datos de la mejor manera posible. Para determinar si una componente aporta información relevante existen diversos criterios, tales como la cantidad de varianza explicada por dicha componente, el valor propio (se exige ser superior a uno), o bien indicadores de bondad de predicción como el estadístico Q^2, el cual se calcula a partir de técnicas de validación cruzada.

Este último criterio basado en la Q^2 se utiliza en los softwares ProSensus y SIMCA, los cuales se han empleado en este trabajo. De forma intuitiva se puede decir que el estadístico Q^2 permite evaluar cómo se modifica la bondad de ajuste de una componente principal si parte de las observaciones se mantienen fuera del modelo. Se considera que una componente aporta información relevante si la dirección de variabilidad asociada a dicha componente no se modifica "demasiado" aunque se descarte una parte de las observaciones, en cuyo caso el Q^2 de dicha componente será positivo.

Para identificar observaciones atípicas en el modelo hay que verificar su distancia al modelo por medio del Error Predictivo Cuadrático para el conjunto X de datos (SPE-x). Esto permite identificar países con valores en sus variables que rompen con la estructura de correlación a nivel mundial. Por otra parte, también es conveniente observar el gráfico T^2 de Hotelling, muy empleado en control de calidad para controlar simultáneamente características correlacionadas de un proceso.

La siguiente figura muestra un ejemplo de una observación con alto SPE-x y T^2, así como otra observación con elevado T^2 pero que no dista excesivamente del modelo.

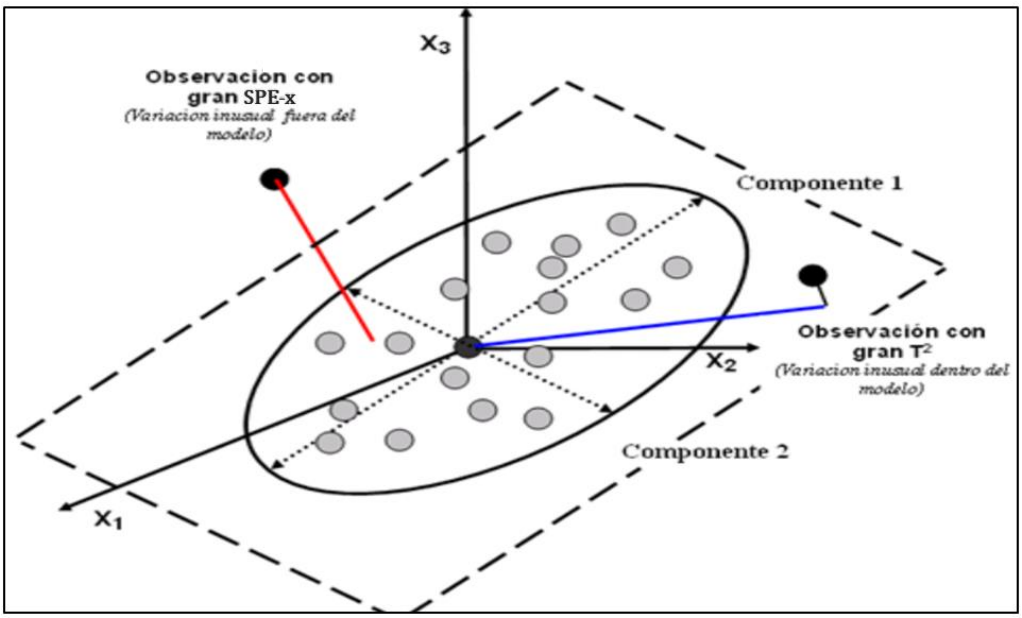

Figura 8. Diferencias entre la distancia al modelo medida por el Error Predictivo Cuadrático (SPE-x) y observaciones atípicas con un elevado T^2 de Hotelling

En la Figura 8 se puede observar en la parte superior izquierda una observación que posee una elevada distancia al modelo (SPE-x). También se muestra una observación atípica ubicada en la derecha de la figura que se encuentra en cercanía al modelo (SPE-x bajo) pero cuyos valores son anormalmente extremos, quedando por fuera de la elipse de Hotelling (alto T^2), aunque se mantiene razonablemente la estructura de correlación de las variables. Este estadístico T^2 de Hotelling se ha empleado en los modelos PCA y PLS.

2.2.2. Regresión de Mínimos Cuadrados Parciales - PLS

La regresión PLS (*Partial Least Squares Regression*) es un método estadístico basado en estructuras latentes cuyo algoritmo tiene cierta relación con el de PCA. Fue desarrollado por Wold (1966a, 1966b) en base a la combinación de modelos de regresión econométricos y el algoritmo NIPALS.

Al ser la regresión PLS un método de proyección sobre estructuras latentes, muchos conceptos son comunes al PCA. También utiliza componentes, que son las direcciones sobre las que se proyectan las observaciones (países), pero en PLS las componentes no son las principales ya que no maximizan la varianza de la matriz X. En caso de una sola variable dependiente Y, el objetivo de la primera componente es obtener una variable latente que cumpla con una situación intermedia entre maximizar la correlación con la variable dependiente y explicar la mayor cantidad posible de varianza de la matriz *"X"*. En concreto es aquella que maximiza la covarianza con la variable respuesta.

La regresión PLS tiene como utilidad modelizar relaciones fundamentales entre las matrices *"X"* e *"Y"*, encontrando variables latentes en el espacio de *"X"* que maximizan la covarianza con las variables latentes del espacio *"Y"*. El modelo PLS funciona bien en el caso de alta multicolinealidad en las variables de *"X"*.

Para entender de mejor manera cómo se construye la primera componente del modelo PCA y PLS, a modo de ejemplo se consideran 6 países (tres de ellos desarrollados y tres no desarrollados) en el espacio de tres variables como se observa en la Figura 9. La primera componente del modelo PCA busca explicar la máxima variabilidad de los datos, mientras que el modelo PLS busca explicar la variable dependiente, en este caso la discriminación entre países en función de su desarrollo.

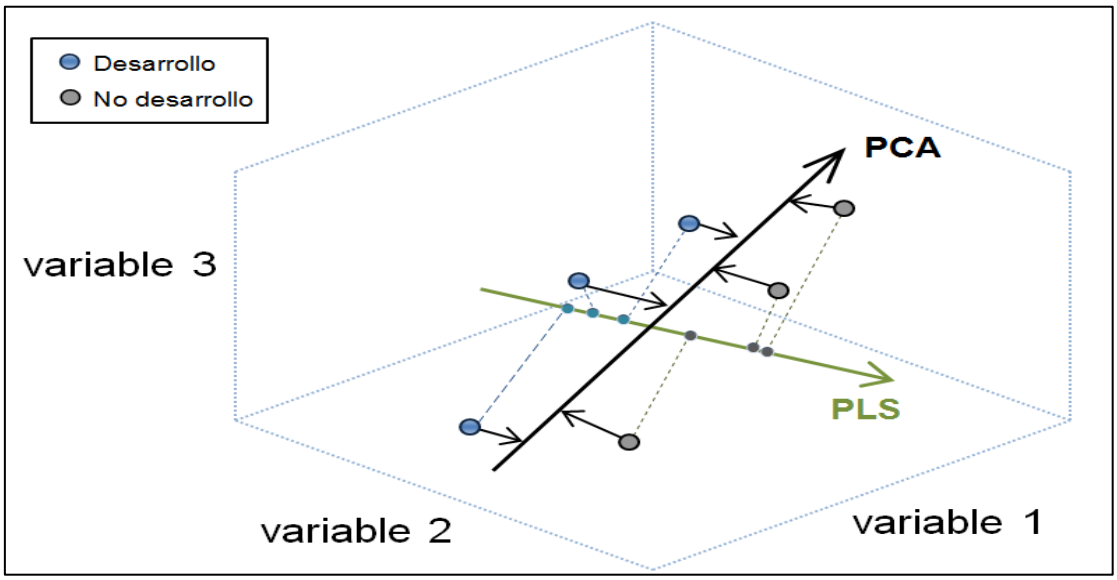

Figura 9. Gráfico donde se ilustra la primera componente PCA (dirección de máxima variabilidad en el espacio de tres variables) así como la primera componente PLS (variable latente que mejor discrimina entre países con desarrollo alto frente a bajo)

2.2.3. Modelo de Regresión por Componentes Principales - PCR

La regresión basada en componentes principales, denominada PCR (*Principal Components Regression*), es un método en dos etapas. En primer lugar se realiza un análisis PCA sobre la matriz X y se determina el número de componentes que aportan información relevante. Se estudia también la presencia de observaciones anómalas que tuvieran que ser descartadas del estudio. A continuación se extraen las variables latentes asociadas a las componentes principales relevantes (vectores de *scores*), y posteriormente con estas nuevas variables se construye un modelo de regresión lineal múltiple para cada una de las variables respuesta. Este modelo puede ajustarse empleando la regresión paso a paso (*stepwise*) en el caso de que el número de variables explicativas sea elevado. La significación estadística de las variables latentes en el modelo de regresión se evalúa a partir del p-valor asociado a cada variable. Finalmente, el modelo se valida estudiando la normalidad de los residuos y verificando la posible existencia de efectos cuadráticos.

2.2.4. Tratamiento de datos faltantes en los modelos PCA y PLS

El análisis PCA se lleva a cabo por medio del algoritmo NIPALS (*Nonlinear Iterative Partial Least Squares*), el cual está implementado en el software estadístico Prosensus y SIMCA-P empleado en este trabajo. Dicho algoritmo permite manejar cierta cantidad de datos faltantes, lo cual es una ventaja ya que en la práctica es habitual encontrarse con este problema. El algoritmo NIPALS es un método secuencial mediante el cual se calcula de forma iterativa las distintas componentes principales de una matriz de datos (Geladi y Kowalski, 1986).

La regresión PLS ejecutada con ProSensus y SIMCA-P también permite trabajar con una cierta cantidad de datos faltantes ya que utiliza el algoritmo NIPALS, el cual de forma iterativa calcula componentes del modelo, así aprovecha los datos conocidos de las observaciones faltantes y los proyecta en el modelo como si cayeran sobre la recta de regresión del modelo y así se completa la base de datos sin alterar dicho modelo.

3. RESULTADOS DE LOS MODELOS ESTADÍSTICOS

3.1. Análisis exploratorio

En un contexto de globalización en el cual se encuentra inmerso el mundo actual, es importante entender las relaciones entre países desde un punto de vista multidimensional. De hecho, los países hoy en día tienden a conformar grupos o asociaciones en base a intereses políticos, económicos, ambientales, productivos, etc.

Los intereses en común entre países permiten estas asociaciones. Pero estos intereses están relacionados muchas veces con la "afinidad" o "similitud" entre naciones. Así pues, los países tienen similitudes en diversos ámbitos como culturales, sociales o poblacionales que les permiten asociarse, parecerse y diferenciarse entre sí.

Este trabajo busca estudiar las similitudes y diferencias entre países. Para ello se ha empleado como modelo exploratorio el Análisis de Componentes Principales (PCA). Esta técnica, como se ha explicado en la sección 2.2.1, permite observar similitudes y diferencias entre los distintos países del mundo. Además, permite reducir la dimensionalidad de los múltiples indicadores, obteniendo estructuras latentes a partir de las cuales se pueden mostrar gráficamente las similitudes y diferencias entre países y regiones geográficas.

El modelo PCA también permite analizar las variables que hacen que los países se parezcan o se diferencien entre sí, mostrando así agrupaciones de variables con relación proporcional directa o indirecta, o bien variables que no presentan relación alguna. Este análisis puede ayudar a entender las estructuras de correlación entre los múltiples indicadores empleados en este estudio. Una mejor comprensión de estas relaciones resulta útil para desarrollar o fortalecer políticas públicas, a nivel nacional o internacional, que reduzcan la diferencia entre países desarrollados con alta calidad de vida frente a los vulnerables y subdesarrollados.

Con el fin de analizar la relación entre países y de observar la dinámica en el transcurso del tiempo de las variables de forma exploratoria, el despliegue más conveniente de la matriz tridireccional es mantener los 139 países y considerar 48 x 14 variables. Es decir, se considera un despliegue de los datos tal como se puede observar en el siguiente esquema:

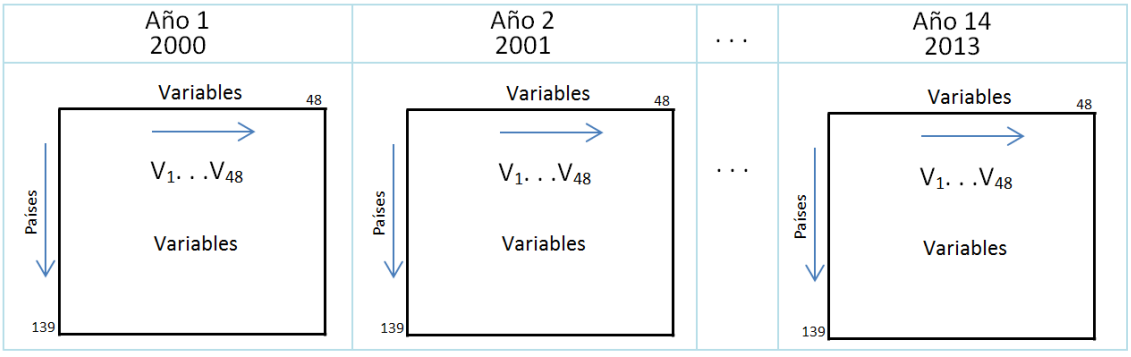

Figura 10. Despliegue de la matriz X tridireccional, resultando 139 países (en filas) por 672 variables (14años *48 indicadores)

3.1.1. Resultados del modelo PCA

Previamente al PCA generalmente es necesario realizar un pretratamiento de centrado. Consiste en restar a cada dato el valor medio de la variable, de modo que todas las nuevas variables centradas tendrán media cero. En este estudio, algunas variables tienen un rango amplio de variabilidad (ingreso per cápita, número de habitantes, número de muertes por SIDA, etc.), mientras que otras variables tienen un rango comprendido entre 0 y 1 (coeficiente de Gini, indicadores de porcentajes, etc.). Cuando se presentan estas situaciones, conviene escalar todas las variables a varianza unitaria antes de ejecutar el PCA, dividiendo cada dato por la desviación estándar de la variable. Con esto se consigue que todas las variables tengan varianza unitaria. El pretratamiento de centrado y escalado a varianza unitaria se suele denominar "autoescalado" por ser la opción más común en análisis multivariante, la cual está implementada por defecto en muchos programas informáticos para PCA.

Ya que los modelos PCA y PLS se ven influenciados notablemente por la escala de las variables, al utilizar este tratamiento (variables autoescaladas) se evita un posible sesgo por esta diferencia de rangos. Por otra parte, permite comparar variables de diferentes unidades y controlar *a priori* el peso de cada variable en los modelos.

La cantidad de varianza que explica cada componente se denomina bondad de ajuste (R^2). El software empleado, tanto Prosensus como SIMCA-P, calcula también este estadístico por técnicas de validación cruzada, lo que se denomina Q^2. Este parámetro resulta útil para decidir el número de componentes del modelo PCA que conviene considerar, y que sintetizar la información relevante de la matriz X.

Para identificar observaciones anómalas hay que revisar la distancia al modelo por medio del Error Predictivo Cuadrático para el conjunto de datos X (SPE-x), lo cual permite identificar países con valores en sus variables que rompen la estructura de correlación de las variables a nivel mundial. Por otra parte, los gráficos basados en la T^2 de Hotelling, ampliamente utilizados en control de calidad, también convienen ser revisados.

3.1.1.1. Modelo PCA inicial con todos los países

Los modelos PCA se han realizado con la matriz original desdoblada (Figura 10) con datos incompletos para un conjunto de 139 países y 48·14 variables. El software empleado se basa en el algoritmo NIPALS que permite manejar cierta cantidad de datos faltantes, como se lo explica en la sección 2.2.4.

La Tabla 3 muestra la bondad de ajuste R^2 de cada componente para este primer modelo, así como la bondad de ajuste obtenida por validación cruzada (Q^2). Los valores indican que la primera componente explica un 41,7% de la variabilidad total de los datos autoescalados, la segunda componente explica un 9% adicional, la tercera explica un 6 % adicional y la cuarta, un 5% adicional.

Tabla 3. Características de las ocho componentes principales del modelo PCA inicial para todos los países

Componentes	R^2_X	R^2_X (acum.)	Valor propio	Q^2	$Q^2_{límite}$	Q^2 (acum.)
1	0,417	0,417	20	0,397	0,0209	0,397
2	0,091	0,508	4,38	0,067	0,0213	0,437
3	0,058	0,566	2,76	0,038	0,0218	0,459
4	0,053	0,619	2,53	0,038	0,0222	0,479
5	0,039	0,658	1,88	0,015	0,0227	0,487
6	0,036	0,694	1,72	-0,010	0,0232	0,482
7	0,027	0,721	1,3	-0,014	0,0238	0,475
8	0,025	0,746	1,22	-0,028	0,0243	0,460

El estadístico Q^2 se obtiene por validación cruzada. El aumento innecesario del número de componentes provoca sobreajuste, el cual no resulta deseado. En general, se considera que una componente aporta información relevante cuando su Q^2 es superior a cero, lo cual implica que la componente tiene capacidad para modelizar la relación entre variables empleando un algoritmo de validación cruzada. Los datos de la tabla 3 se han obtenido con el software SIMCA-P 10.0. Este software es un poco más restrictivo, y considera como relevantes aquellas componentes cuyo Q^2 supera cierto valor umbral denominado Q^2 límite. Este valor umbral no se considera en el programa ProSensus, lo cual significa que este software implícitamente asume un valor umbral nulo.

El programa Prosensus también calcula el estadístico Q^2 para cada componente por medio de un método de validación cruzada. Dicho algoritmo para calcular Q^2 puede ser ligeramente distinto entre programas. No obstante, se ha ejecutado el mismo modelo con ProSensus y SIMCA-P, y se ha comprobado que los valores de Q^2 son bastante parecidos.

A la vista de la tabla 3, se concluye que las cuatro primeras componentes principales son las que modelizan la información relevante (Q^2 positivo y superior al valor umbral). Con dichas componentes se explica el 53% de la variabilidad de los datos.

Para validar de forma global el modelo es necesario inspeccionar tanto la distancia de las observaciones al modelo a través del error cuadrático de predicción (SPE-X), como el gráfico de la T^2 de Hotelling.

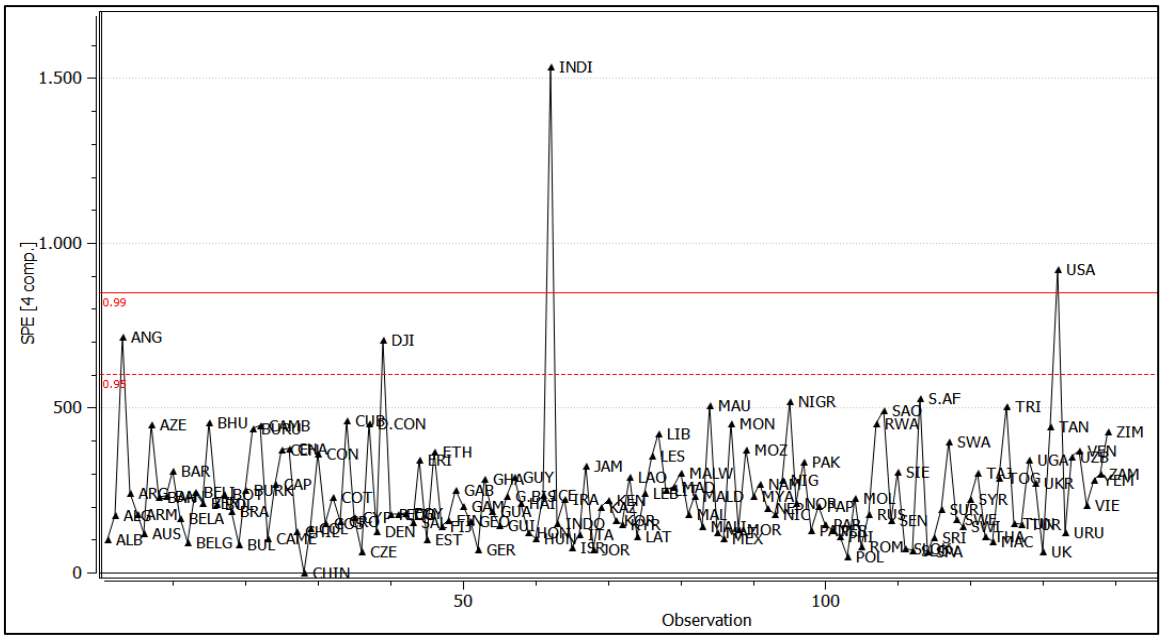

Figura 11. Error cuadrático de la predicción del primer modelo PCA con cuatro componentes

En el gráfico de SPE-X se observa a India y Estados Unidos como países atípicos. La Figura 12 muestra el valor del estadístico T^2 de Hotelling considerando cuatro componentes. Se observa que China es claramente un país atípico, el cual se ha comprobado que ejerce una influencia excesiva sobre la segunda componente principal (Figura 13).

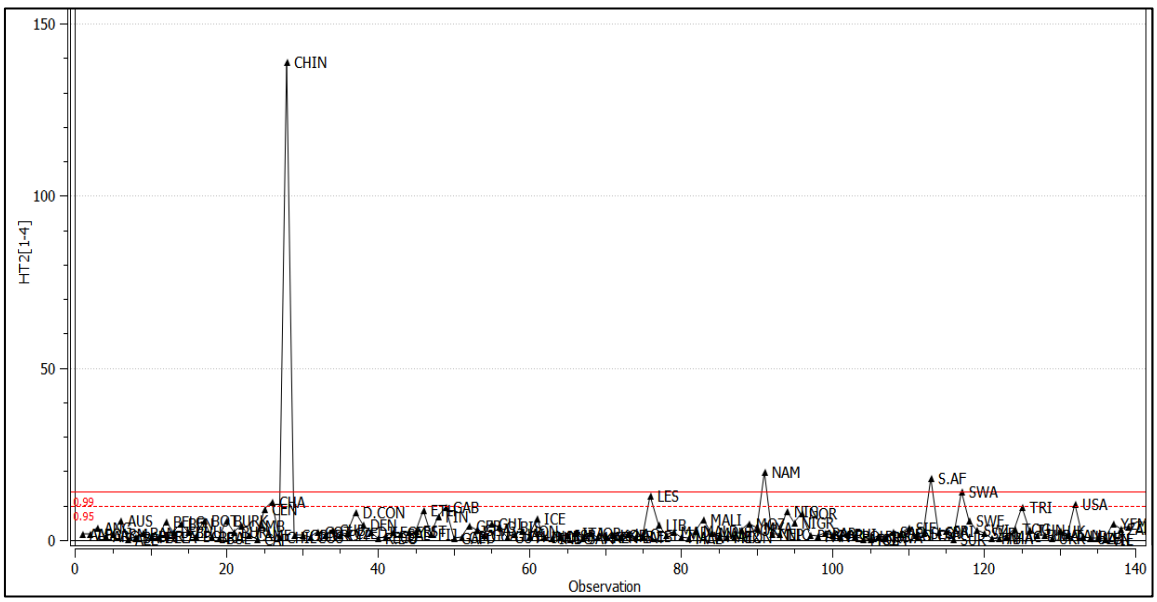

Figura 12. T^2 de Hotelling del modelo inicial (componente 1 a la 4)

Este comportamiento atípico de China observado en las figuras 12 y 13 con respecto a los demás países, indica que este país debe ser estudiado por separado, pues tiene valores no comparables con el resto de países del mundo. La Figura 13 revela que el modelo PCA puede estar bastante influenciado por China, de modo que este país debe descartarse para estudiar convenientemente las estructuras de correlación del resto de países.

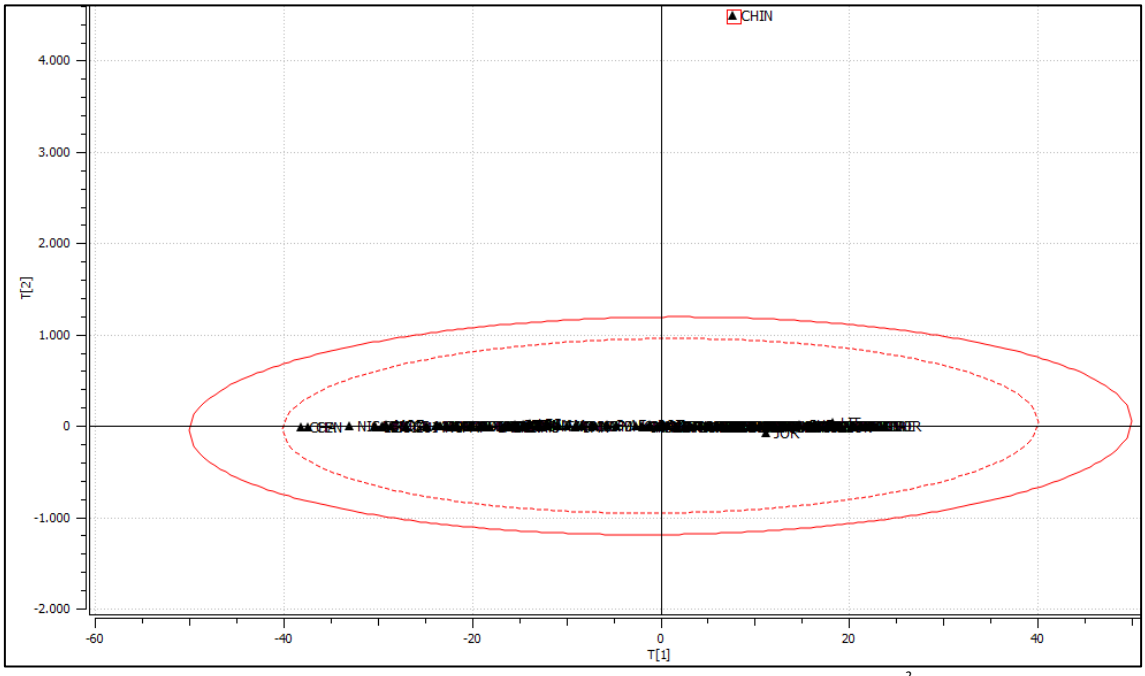

Figura 13. Gráfico de scores (T2 frente a T1) del modelo inicial, donde se observa la elipse T^2 de Hotelling

Para caracterizar las diferencias de China con respecto al promedio mundial, en la Figura 14 se muestra la contribución de China respecto a la media de países en el espacio de las X, en la cual se observa que las variables "CO_2 per cápita", "consumo de sustancias que agotan al ozono" y "población total" tienen valores anormalmente elevados.

Los resultados muestran a China como una observación influyente en el modelo y por este motivo se decide eliminar a este país del resto del estudio.

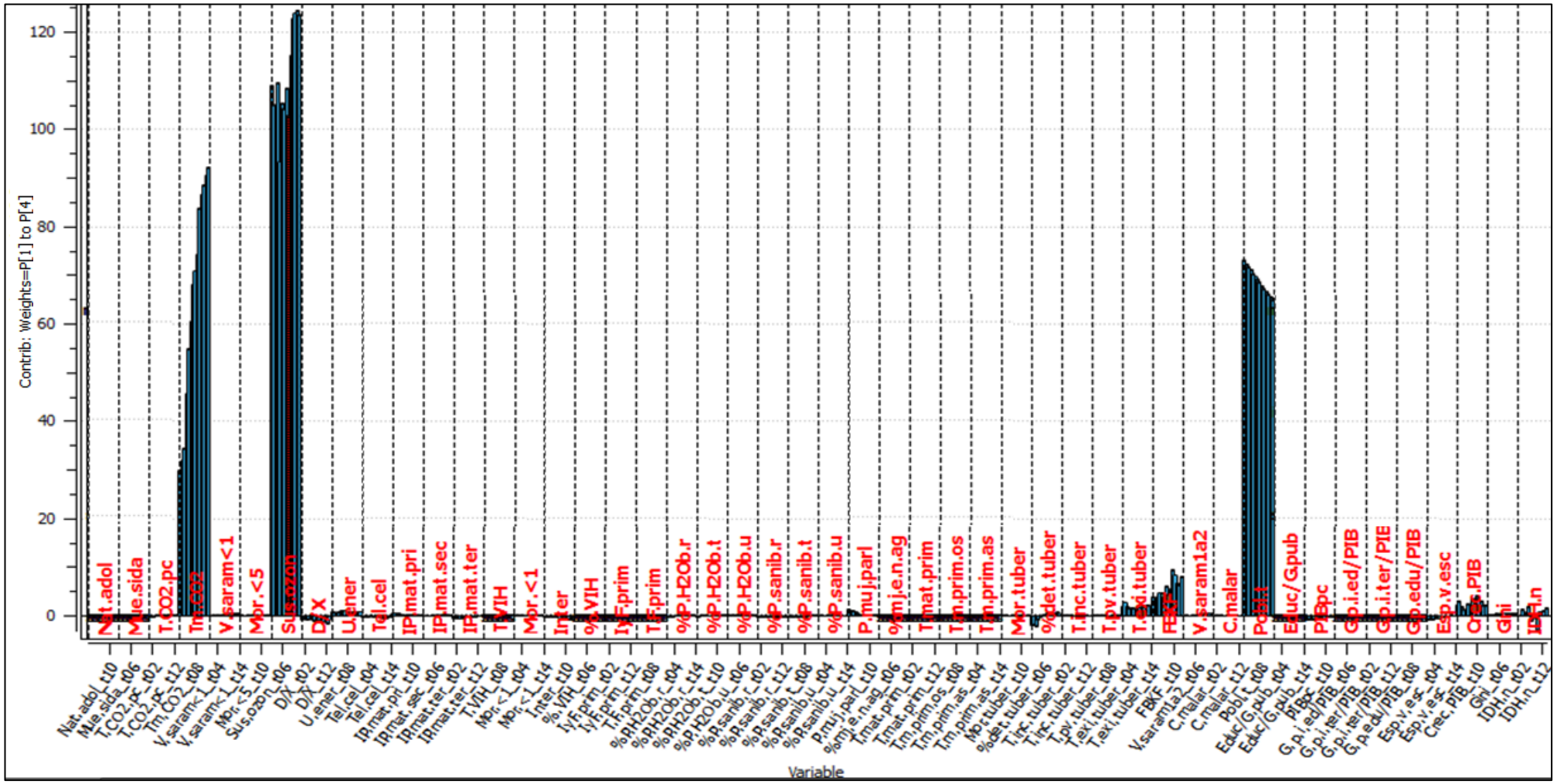

Figura 14. Contribución de China respecto a la media de países al espacio de las X, componentes de la 1 a la 4

3.1.1.2. Modelo PCA sin China

De este segundo modelo con 138 países, tras descartar a China, se observa en la tabla 4 que las cinco primeras componentes satisfacen el criterio de Q^2 positivo y por tanto aportan información relevante. Los valores de R^2 son muy similares a los de la Tabla 3. Por este motivo, se considerará un modelo con cinco componentes.

Tabla 4. Características de las ocho componentes principales del modelo PCA sin China

Componentes	R^2_X	R^2_X (acum.)	Valor propio	Q^2	$Q^2_{límite}$	Q^2 (acum.)
1	0,417	0,417	20	0,397	0,0209	0,397
2	0,091	0,509	4,2	0,067	0,0213	0,437
3	0,058	0,567	2,72	0,023	0,0218	0,459
4	0,053	0,619	2,5	0,0017	0,0222	0,479
5	0,039	0,659	1,78	0,015	0,0227	0,487
6	0,036	0,694	1,7	-0,009	0,0232	0,482
7	0,027	0,722	1,33	-0,014	0,0238	0,475
8	0,025	0,747	1,25	-0,027	0,0243	0,460

Observando el error cuadrático de la predicción de este modelo se deduce que no hay ningún país con una considerable distancia al modelo que requiera ser descartado.

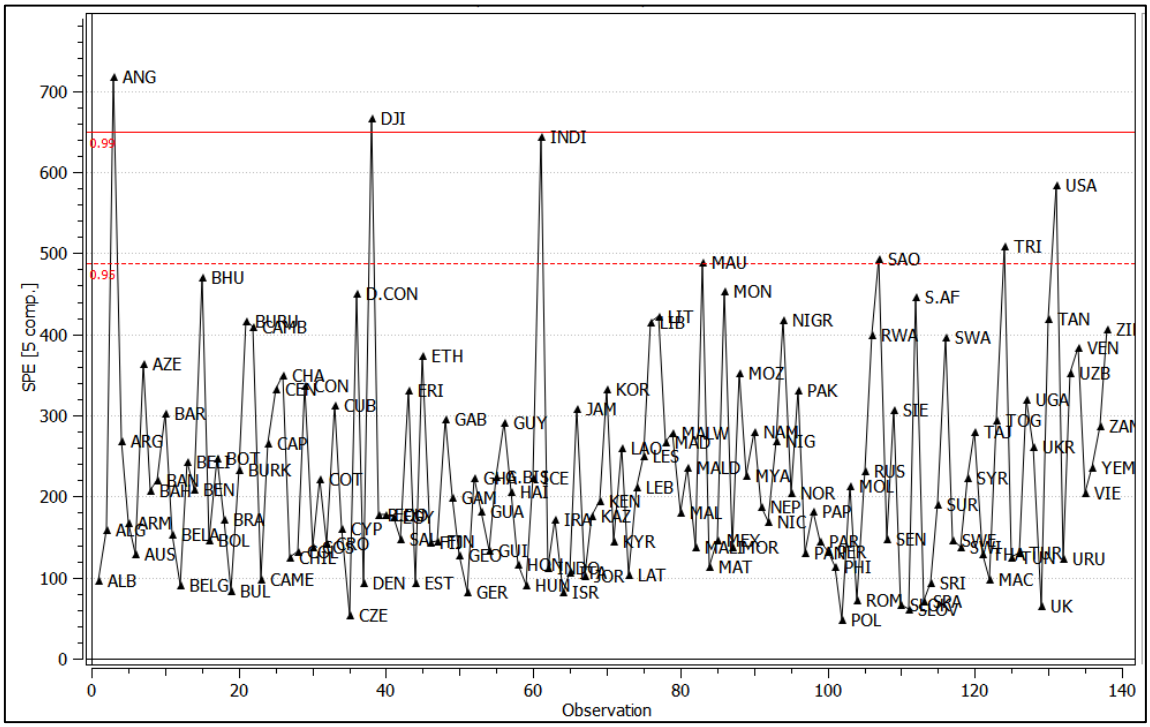

Figura 15. SPE en X - Modelo PCA sin China con cinco componentes

En la Figura 16 se muestra la T^2 calculada con 5 componentes. Se observa claramente que India y Estados Unidos sobresalen de los niveles del 99%. Ambos países sobrepasan el umbral del 95% en el gráfico SPE-X (Figura 15). Así pues, dada la clara anomalía de estos dos países, han sido también descartados del resto del estudio.

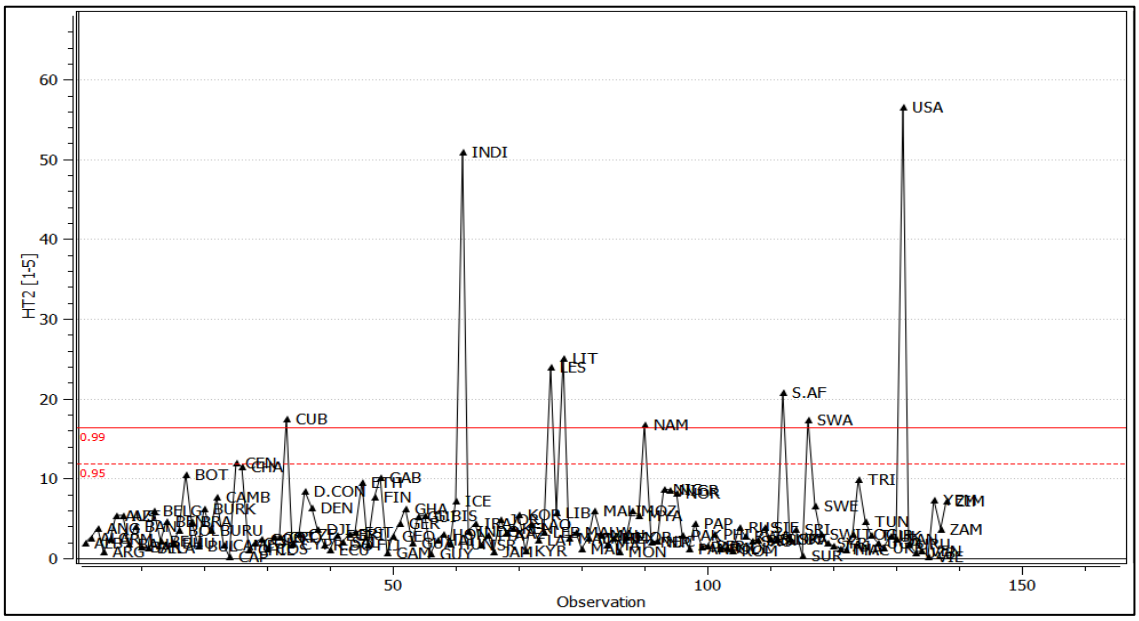

Figura 16. Gráfico T^2 de Hotelling calculado con 5 componentes (modelo PCA sin China)

Por otra parte, estudiando el gráfico de contribución a los *scores* de India respecto al promedio de las observaciones, se ha observado que las variables de mayor contribución son "población total" y "casos de malaria", lo cual indica que India tiene valores muy elevados de ambas variables. En el caso de Estados Unidos, se ha comprobado que este país tiene altos valores de emisión de CO_2 per cápita.

Así pues, se decide excluir del modelo final a los tres países atípicos: China, India y Estados Unidos.

3.1.1.3. Modelo PCA sin China, India y USA

Tras descartar estos tres países (quedan 136 en el modelo) se ha repetido el PCA y, curiosamente, el número de componentes relevantes, con Q^2 positivo, pasa a ser básicamente de dos o quizás tres (tabla 5). El estadístico de bondad de ajuste R^2 indica que la primera componente explica un 41,8% de la variabilidad, la segunda explica un 8,4% adicional, la tercera un 5,5 % adicional y la cuarta otro 5%. Estos valores son casi idénticos a los dos modelos PCA anteriores. El modelo con dos componentes explica acumuladamente un 50,3% de la variabilidad total de los datos.

Tabla 5. Características de las ocho componentes principales del modelo PCA sin China, India y USA

Componentes	R^2_X	R^2_X (acum.)	Valor propio	Q^2	$Q^2_{límite}$	Q^2 (acum.)
1	0,418	0,418	20,1	0,398	0,0209	0,398
2	0,084	0,503	4,05	0,089	0,0213	0,452
3	0,055	0,558	2,66	-0,011	0,0218	0,446
4	0,050	0,609	2,43	0,026	0,0223	0,460
5	0,040	0,649	1,95	0,002	0,0227	0,462
6	0,037	0,687	1,8	0,012	0,0232	0,468
7	0,026	0,713	1,28	-0,010	0,0238	0,463
8	0,025	0,739	1,21	-0,023	0,0243	0,450

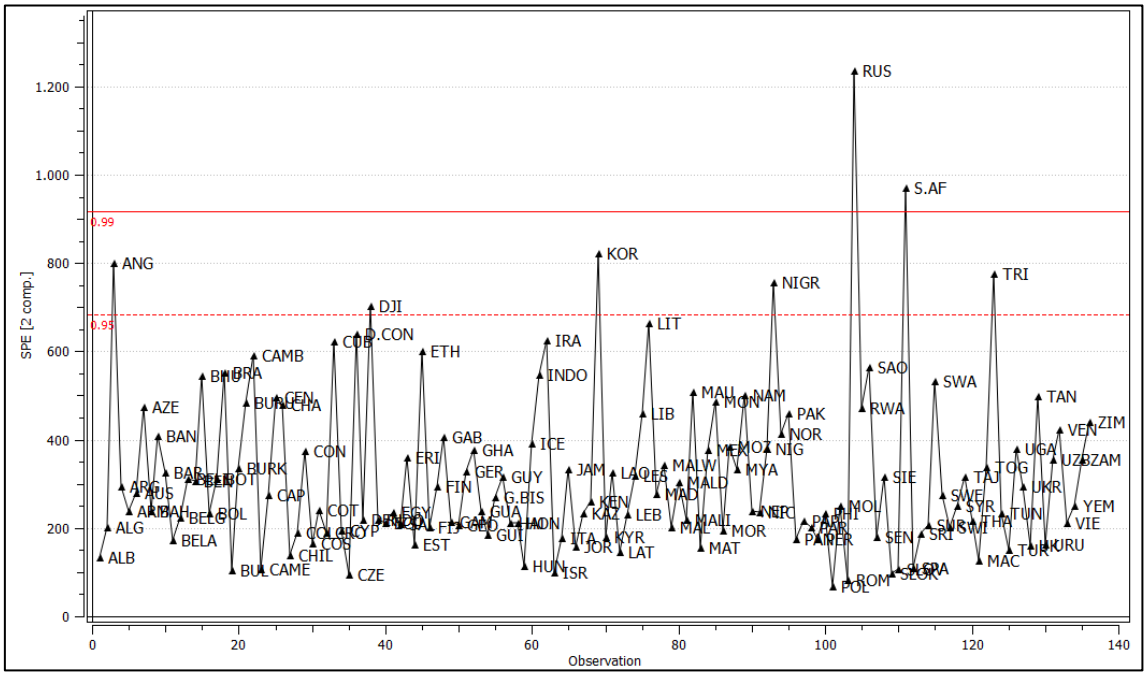

Figura 17. SPE en X del modelo PCA sin China, India y Estados Unidos (calculado con dos componentes)

A partir del gráfico SPE-X (figura 17) se observa que Rusia y Sudáfrica (S.AF) presentan un valor de este estadístico SPE-X ligeramente por encima del nivel de confianza del 99%. Esto implica un leve comportamiento atípico, por lo que se ha decidido mantener ambos países en el modelo. En el caso de Rusia, la anomalía parece ser que está causada por un dato atípico en la serie de consumo de sustancias que agotan el ozono (año 2000) que no es atípico en los años siguientes. Este valor elevado se debe a que Rusia a partir del año 2000 implementó políticas para reducir el consumo de sustancias que agotan al ozono, por lo que hay un cambio en la tendencia de la variable desde el año 2001.

3.1.2. Análisis de similitudes y diferencias entre países

Los resultados del PCA permiten discutir las similitudes y diferencias entre los países. En el gráfico de *scores*, los países que se encuentran en cercanía euclidiana tienen valores similares en sus variables. Esto permite observar similitudes con sus "vecinos" en ámbitos sanitarios, sociales o políticos, entre otros.

Con el objetivo de facilitar la visualización de resultados se han coloreado los países en función de su región geográfica (la cual no se contempla como variable en el modelo PCA). Las elipses dibujadas en negro corresponden a países con cierta similitud con sus vecinos geográficos. Estas similitudes y diferencias se reflejan en la figura 18.

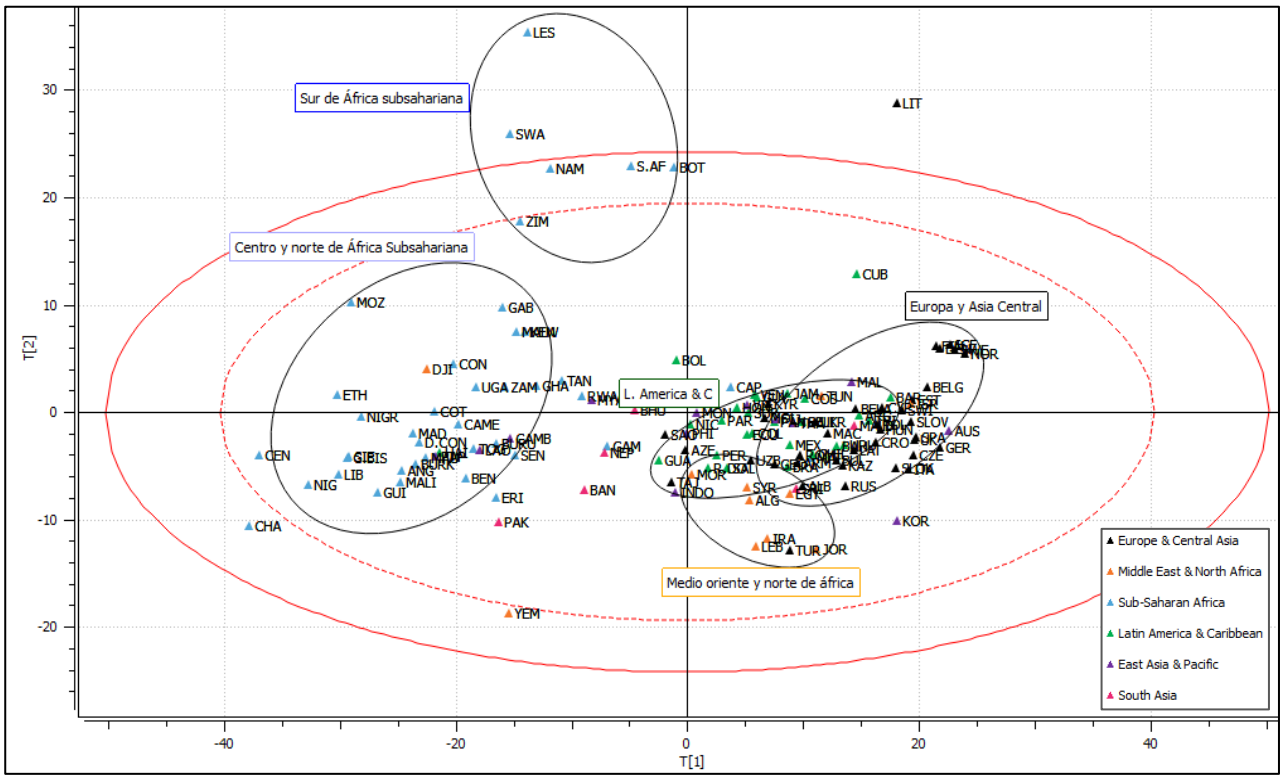

Figura 18. Gráfico de scores (T_2 frente a T_1) del modelo PCA sin China, India y USA

El gráfico de los *scores* (componentes T_1 y T_2) muestra que la primera básicamente discrimina entre los países de África Subsahariana frente al resto. Además se observa una clara separación de algunos países del sur del África Subsahariana respecto al resto del mundo. Esta separación queda reflejada por la componente segunda. Dado que estos países muestran en general una distinción significativa respecto al resto, se realiza más adelante un análisis por separado de este grupo de naciones.

Respecto a los países de Europa comparados con los de Asia Central, no se muestra una división significativa en cuanto a estas dos componentes. Lo mismo sucede comparando los países de Medio Oriente frente al Norte de África. Análogamente, América Latina y Caribe aparecen sin distinción.

En particular, se observa una separación de cinco países escandinavos (Noruega, Islandia Finlandia, Suecia, Dinamarca y Bélgica) con respecto a los países de la región. Estos cinco países muestran similitudes en valores bajos de mortalidad en niños menores de 5 años, altos valores de PIB per cápita, gran cantidad de usuarios de internet por 100 habitantes, y una alta emisión de CO_2 per cápita.

Se observa a Lituania (LIT) como un caso especial que aparece fuera de los límites de la elipse T^2 de Hotelling y muy separado de su región (Europa y Asia Central). La anomalía de Lituania se debe a que presenta altos valores para las siguientes variables: “gasto público en educación/PIB”, “gasto público en instituciones educativas/PIB” y “gasto público en instituciones de tercer nivel”, las cuales muestran una distinta estructura de correlación de las variables respecto al resto del mundo.

Curiosamente, Lituania presenta un valor de T^2 similar a Botswana (BOT) y Lesoto (LES) en la figura 18 debido a sus altos valores en gasto público en educación/PIB, pero sus proyecciones sobre la primera componente (relacionadas con incidencia de enfermedades) son muy diferentes.

Respecto a la región América Latina y Caribe, se observa que Cuba (CUB) muestra una separación con el resto de países (Figura 18). Esto es debido a que Cuba presenta altos valores en tres variables asociadas a gasto público en instituciones de educación, educación de nivel terciario y gasto en educación en general, referidas en porcentaje respecto al PIB.

Se observa a Trinidad y Tobago como un caso curioso, ya que se proyecta cercano a países considerados económicamente desarrollados. El motivo parece ser que muestra similitud sólo en cuanto a altos valores de emisión de CO_2 per cápita.

Para la región de Medio Oriente y Norte de África se observa que Yemen (YEM) se aleja de los demás países (posición inferior en la Figura 18). Esto es debido a que Yemen muestra valores muy bajos en tres variables asociadas a igualdad de género: “porción de mujeres entre los empleados remunerados en el sector no agrícola”, “índice de paridad de matrícula primaria” e “índice de paridad de matrícula secundaria”.

Respecto a la región de América Latina y Caribe, resulta curioso observar que Haití muestra similitudes con los países de África Subsahariana. Esto se debe a sus valores bajos en porcentaje de población que utiliza fuentes de agua mejoradas, bajo porcentaje de población que utiliza servicios mejorados de saneamiento, baja tasa de vacunación contra sarampión en niños menores de 2 años y valores altos de mortalidad infantil.

Para comprender mejor las diferencias entre África Subsahariana respecto a las demás regiones del mundo, en el siguiente gráfico (figura 19) se muestra la contribución de un país elegido al azar de esta zona (Nigeria) en comparación con el promedio mundial. Esta gráfica permite identificar las diferencias de los países de esta zona (Norte, Sur y Centro de África Subsahariana) respecto al resto del mundo.

Entre las variables de mayor diferencia aparecen indicadores relacionados con i) elevada mortalidad por enfermedades causadas por tuberculosis y sida, ii) vulnerabilidad infantil como “mortalidad de niños menores a un año”, bajo “porcentaje de niños de un año y niños entre uno y dos años vacunados contra el sarampión” y baja “tasa de matrícula primaria”, y iii) calidad de vida, con una baja “proporción de la población que utiliza fuentes mejoradas de agua potable” y baja “proporción de la población que utiliza servicios de saneamiento mejorados”.

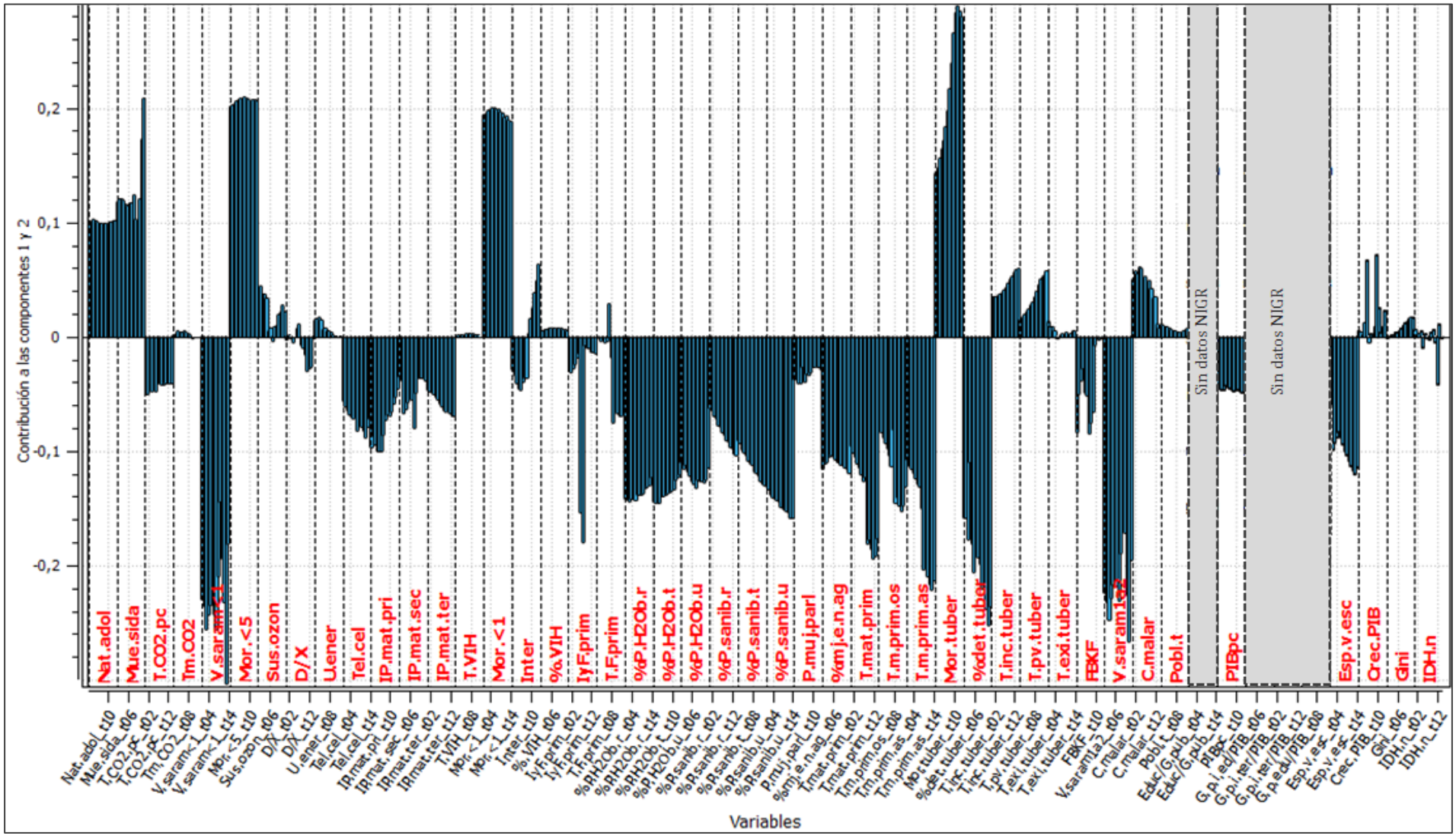

Figura 19. Gráfico de contribución de Nigeria respecto al promedio mundial (modelo PCA con dos componentes)

3.1.2.1. Análisis particular de países del Sur de África

A partir de la Figura 18 se deduce que hay cinco países del sur de África Subsahariana con una elevada influencia en la segunda componente principal. Esto indica que muestran una relación de semejanza entre sí pero que difieren del resto del mundo. Dichos países son Suazilandia, Namibia, Botswana, Lesoto y Sudáfrica. Este grupo, además de la cercanía geográfica (sur de África), presenta valores muy altos en el porcentaje de personas con VIH y tasa de incidencia de VIH, lo cual indica que estos países se encuentran en una situación fuera de control de este virus en relación al resto del mundo. Sudáfrica y Suazilandia se caracterizan además por tener altas tasas de incidencia de tuberculosis.

En el caso de Botswana y Lesoto, cabe mencionar un alto gasto público en educación calculado como porcentaje del PIB. Estas relaciones se pueden observar en la siguiente figura.

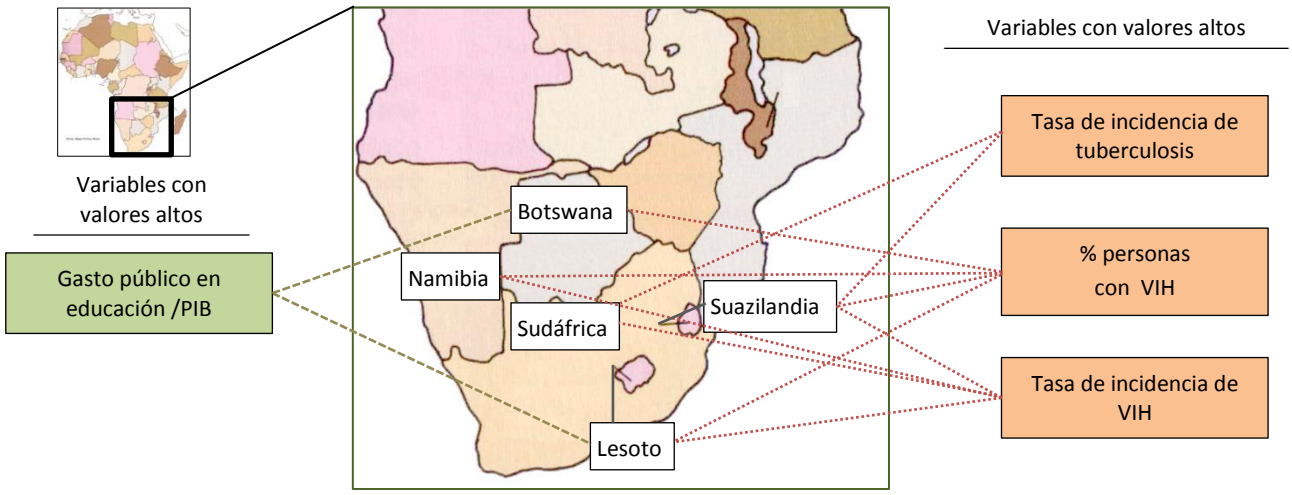

Figura 20. Asociaciones entre los países del Sur de África Subsahariana

3.1.3. Análisis de la relación entre variables a nivel mundial

Los *loadings* resultantes del modelo PCA permiten discutir la relación entre variables a nivel mundial. El gráfico de *loadings* tiene una correspondencia directa con el de *scores*, pudiéndose identificar las variables que más discriminan entre los *clústers* de países discutidos anteriormente. Variables situadas cerca entre sí en el gráfico de *loadings* presentan en general correlación positiva, mientras que una posición en polos opuestos suele corresponder a las relaciones inversas (correlación negativa).

La figura 21 muestra los *loadings* correspondientes a las dos primeras componentes principales (P2 frente a P1). Se observa una relación directamente proporcional entre las variables de mortalidad de niños menores de 5 años y la tasa de madres adolescentes. Estas variables se muestran inversamente proporcionales al grupo de indicadores característicos de los países desarrollados (esperanza de vida escolar, tasa de matrícula primaria, índice de paridad de matrícula a nivel de primaria, secundaria y tercer nivel, acceso a internet, PIB per cápita, vacunas contra sarampión, porcentaje de población con buen servicio de saneamiento y aguas mejoradas, etc.).

El análisis de la relación entre variables es importante sobre todo en un ámbito de política pública y asistencia internacional. Por ejemplo, si se desea reducir la mortalidad infantil se

deberían implementar políticas que fomenten la educación y reduzcan el porcentaje de madres adolescentes. Lógicamente, la mortalidad infantil está relacionada de forma inversa con variables como esperanza de vida escolar, tasa de matrícula primaria, igualdad de género en educación y otras variables asociadas al desarrollo humano y buena calidad de vida.

Desarrollar políticas de educación y salud para reducir el porcentaje de madres adolescentes permitiría previsiblemente, además de reducir la mortalidad infantil y aumentar la esperanza de vida escolar, mejorar el índice de paridad (proporción de mujeres con respecto a hombres) de educación a nivel primario, secundario y terciario. Esto permitiría alcanzar "adecuados" valores de los indicadores relacionados con el desarrollo humano y buena calidad de vida.

En base al posicionamiento de las variables, la segunda componente muestra una relación directa entre enfermedades de tuberculosis y VIH. También muestra correlación positiva entre las variables relacionadas con gasto en educación relativo al PIB, como "gasto público en educación total", "gasto público en instituciones de educación" y "gasto público en instituciones terciarias".

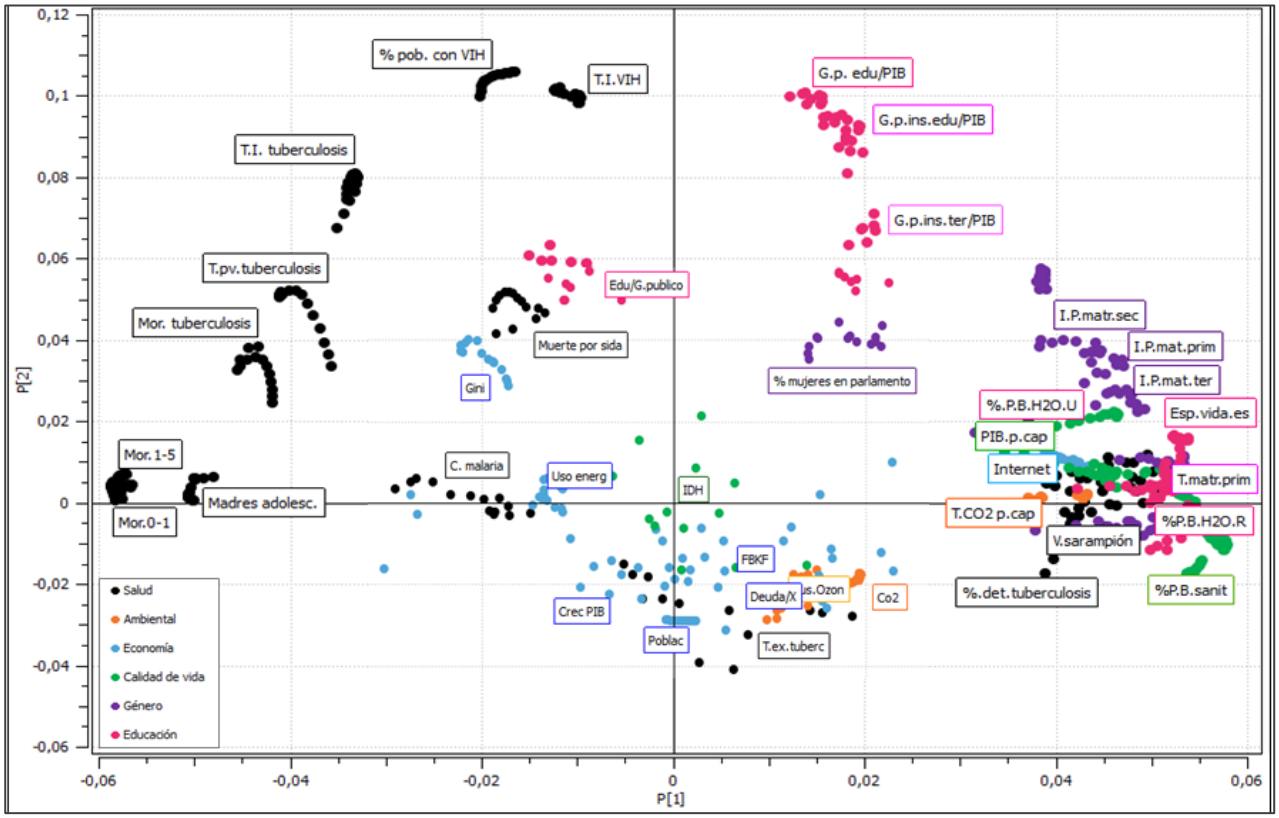

Figura 21. Gráfico de *loadings* (P2 frente a P1) del modelo PCA sin China, India y USA

Se observa que la variable Coeficiente de Gini (que define la distribución equitativa de los ingresos en un país) está muy próxima a "muerte por sida" en la Figura 21, lo cual resulta llamativo. Esto parece sugerir que los países con alta tasa de mortalidad por sida también tienen una distribución equitativa de los ingresos, sean estos altos, medios o bajos. No obstante, esta interpretación es incierta y requeriría un estudio con mayor profundidad.

Las variables en color morado en la Figura 21, asociadas a la igualdad de género (índice de paridad en matrícula primaria, secundaria y terciaria) se encuentran relacionadas con esperanza de vida escolar, tasa de matrícula primaria para ambos sexos y con porcentaje de población que cuenta con fuente de agua mejorada.

Sin embargo, parece ser que esta relación de paridad de género no está correlacionada con un gasto en educación calculado respecto al PIB, pues ambas variables aparecen en posición ortogonal en la figura. Por ello, un mayor o menor gasto en educación no implica necesariamente cambios en la relación de paridad de género en educación.

Los puntos coloreados en azul indican una asociación entre indicadores como "porcentaje de población", "tasa de éxito en tratamiento de tuberculosis", "Formación Bruta de Capital Fijo – FBKF (inversión de un país)", "crecimiento del PIB", "deuda total respecto a exportaciones de bienes y servicios", "consumo de sustancias que agotan al ozono" y "emisión de CO_2". Además, éstas a su vez se encuentran relacionadas de forma inversa con el "coeficiente de Gini", muerte por sida y enfermedades relacionadas con la tuberculosis.

Ambos grupos de variables son ortogonales a los indicadores de mayor peso en la segunda componente: "tasa de incidencia de VIH" y "porcentaje de personas que viven con VIH", los cuales lógicamente están relacionados entre sí.

En el origen de coordinadas de la Figura 21 se encuentra la variable Índice de Desarrollo Humano (IDH), la cual no es explicada en gran medida por ninguna de las dos componentes del modelo PCA. Por ello, para entender la relación de las variables en función del desarrollo, se ha llevado a cabo en la sección 3.2 diferentes modelos de regresión PCR y PLS sobre esta variable.

El gráfico VIP (*Variable Importance Plot*) del modelo PCA, obtenido con dos componentes, muestra la importancia de las distintas variables en este modelo. Las variables de mayor importancia son las que aparecen en el extremo izquierdo de la Figura 21. Las variables de mayor importancia son:

- Tasa de mortalidad infantil (0-1 años) por cada 1.000 nacidos vivos.
- Proporción de la población que utiliza servicios de saneamiento mejorados.
- Tasa de mortalidad de niños menores de cinco años por cada 1.000 nacidos vivos.
- Proporción de la población que utiliza fuentes mejoradas de agua potable.
- Esperanza de vida escolar, primario hasta el terciario (años).
- Porcentaje de alumnos que comienzan el primer grado y llegan al último de primaria.
- Tasa de natalidad entre las adolescentes, por cada 1.000 mujeres.
- Tasa total neta de matrícula en la enseñanza primaria.
- Usuarios de internet por cada 100 habitantes.
- Porcentaje de la población entre 15-49 años que vive con VIH.
- Tasa de incidencia de tuberculosis por año por cada 100.000 habitantes.
- PIB per cápita (en $, precios internacionales constantes en 2011).

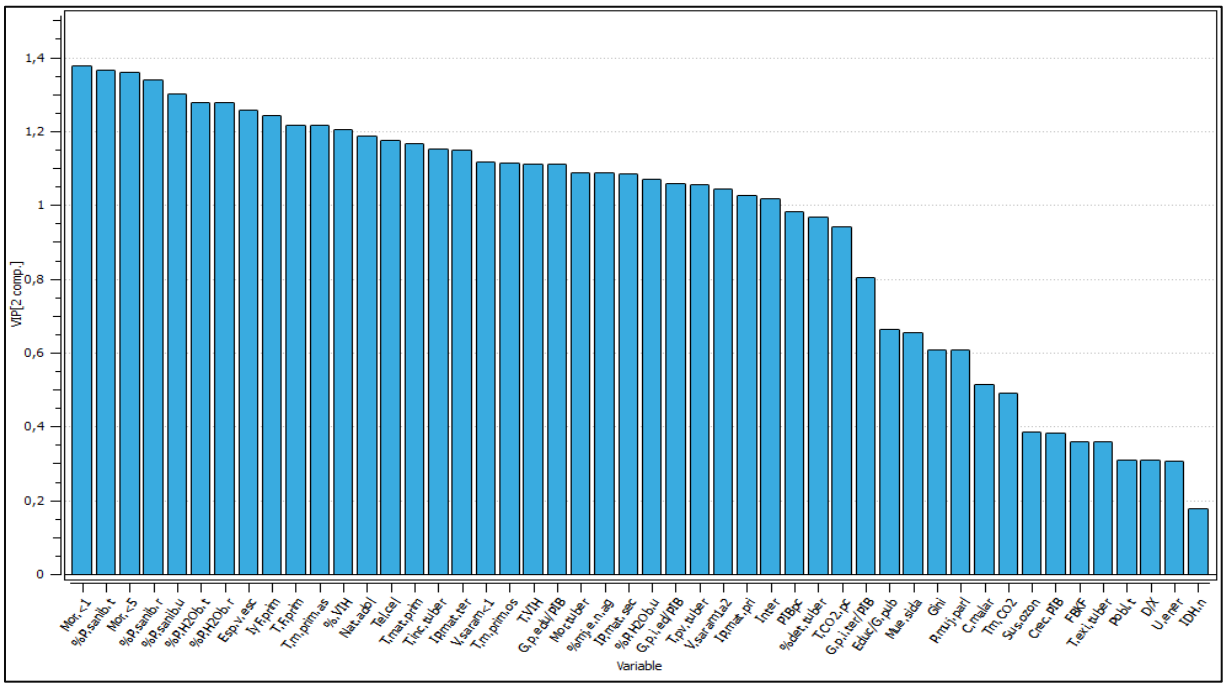

Figura 22. Gráfico VIP de importancia de las variables en el modelo PCA con dos componentes sin China, India y USA

3.1.3.1. Evolución temporal de las variables a nivel mundial

Para evaluar cómo se han modificado los valores de las variables en el período 2000 al 2013, sin considerar a China, India ni Estados Unidos, resulta conveniente realizar un modelo PLS utilizando como variable dependiente "tiempo". Para ello, es necesario un nuevo tipo de desdoblamiento de la matriz. Partiendo de la estructura tridireccional inicial, se desdobla la matriz con 48 variables (columnas) y 139 x 14 observaciones. Posteriormente, se crea una columna denominada "tiempo" que contiene para cada fila de esta nueva matriz desplegada el año al que corresponde la información.

A continuación se ajusta un modelo PLS para predecir esta nueva variable de tiempo. Esta regresión PLS busca identificar qué variables son las que más se han modificado a lo largo del tiempo. El modelo PLS se valida de modo similar al PCA. La tabla 6 muestra los resultados del PLS considerando "tiempo" como variable respuesta. A partir de los valores positivos obtenidos del estadístico Q^2, se deduce que son cinco componentes las que aportan información relevante al modelo, es decir, que tienen capacidad predictiva. En global, las 5 primeras componentes PLS explican el 40,1% de la variabilidad del tiempo.

Tabla 6. Características de las seis componentes PLS considerando Y=tiempo

Componentes	R^2_Y	R^2_Y (acum.)	Q^2	Q^2 (acum.)
1	0,070	0,070	0,062	0,062
2	0,223	0,293	0,183	0,245
3	0,062	0,355	0,054	0,299
4	0,027	0,381	0,031	0,330
5	0,020	0,401	0,032	0,362
6	0,007	0,408	-0,007	0,369

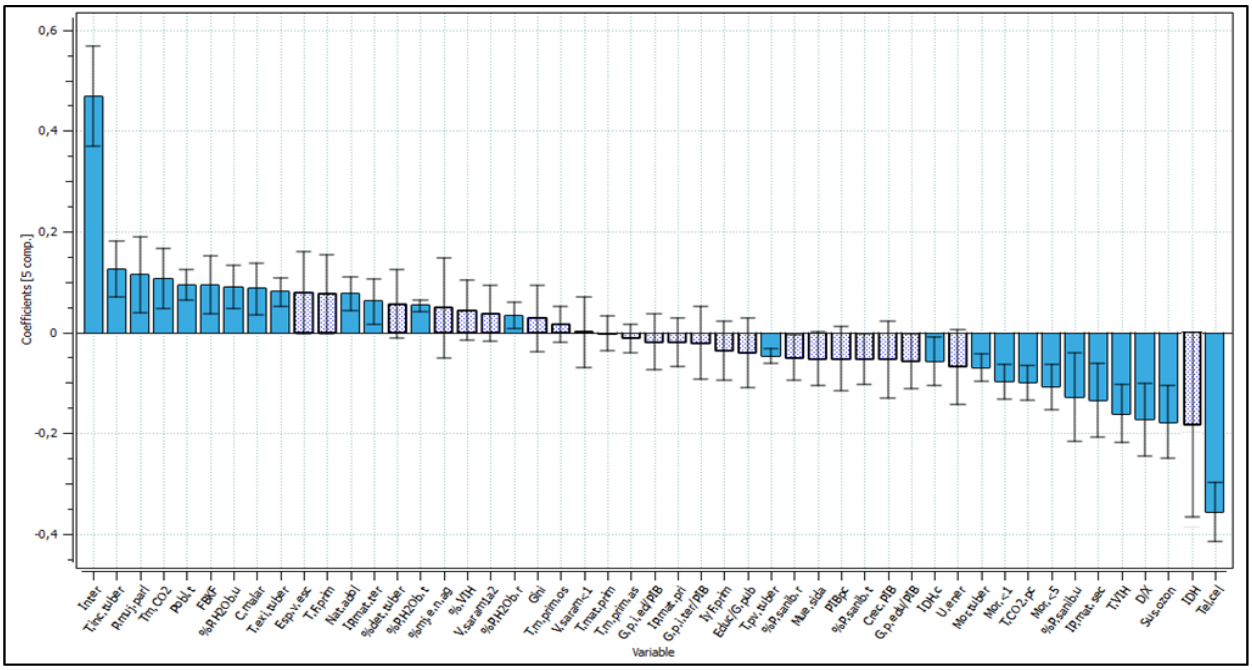

Figura 23. Coeficientes de regresión del modelo PLS con Y=tiempo ajustado con 5 componentes (intervalo de confianza del 95%)

La figura anterior muestra los coeficientes de regresión de este modelo PLS, ajustado con 5 componentes. Este gráfico indica qué variables tienen un coeficiente de regresión significativamente distinto de cero. Esto permite identificar aquellas variables con una tendencia lineal (creciente o decreciente) a lo largo de la serie de años considerada. Es decir, puede deducirse qué variables han mostrado un crecimiento o una disminución en sus valores a nivel mundial. Las variables que muestran una variación estadísticamente significativa son aquellas cuyo intervalo de confianza no contiene el valor 0 (variables de color azul en la figura 23).

Para el período 2000-2013, las variables que han tenido un mayor incremento estadísticamente significativo a nivel mundial son: “acceso a internet”, “tasa de incidencia de tuberculosis por cada cien mil personas”, “puestos ocupados por mujeres en el parlamento”, “emisión total de CO_2”, “población total”, “formación bruta de capital fijo (inversión de un país) respecto al PIB”, “proporción de población que utiliza fuentes mejoradas de agua potable a nivel urbano y nacional”, “casos de malaria reportados”, “tasa de éxito en el tratamiento de la tuberculosis”, “tasa de natalidad entre las adolescentes por cada mil mujeres” e “índice de paridad en matrícula en educación de tercer nivel”.

Las variables que muestran una disminución estadísticamente significativa son: “suscripciones de telefonía fija y celular móviles”, “consumo de todas las sustancias que agotan el ozono”, “servicio de deuda como porcentaje de las exportaciones de bienes y servicios”, “tasa de incidencia del VIH en personas entre 15-49 años”, “índice de paridad de género en la matrícula de nivel secundario”, “proporción de la población que utiliza servicios de saneamiento mejorados a nivel urbano”, “tasa de mortalidad de niños menores de cinco años”, “tasa de mortalidad por tuberculosis” y “tasa de prevalencia de la tuberculosis”.

3.2. Análisis del desarrollo humano

Para analizar el desarrollo humano se han realizado diversos análisis de regresión considerando el Índice de Desarrollo Humano (IDH) como variable respuesta. Dicho índice mide el nivel de desarrollo humano de un país, tal como se ha descrito en la sección 2.1.5. Es un indicador cuantificado con valores estandarizados que oscilan entre 0 y 1.

En este trabajo se han realizado dos modelos de regresión basados en estructuras latentes, que son PLS (regresión de mínimos cuadrados parciales) y PCR (regresión por componentes principales). El objetivo es estudiar qué variables son las que mejor predicen el IDH, lo cual aporta información útil para establecer políticas encaminadas a mejorar el desarrollo de un país. Por otra parte, estos modelos predictivos se han empleado también para predecir el IDH en aquellos países para los cuales este índice no se conoce (42 países), los cuales representan el 18% de todos los 234 países del mundo.

Como se mencionó en la sección 2.2.3, el modelo PCR consiste en una regresión en dos etapas. En primer lugar se ajusta un modelo PCA y se obtienen las variables latentes (vectores de *scores*), y en una segunda etapa se ajusta un modelo de regresión lineal múltiple. Dado que China, Estados Unidos e India son países anómalos, no se han considerado en los modelos PCR ni PLS.

3.2.1. Regresión por Componentes Principales - PCR

La siguiente figura muestra el desdoblamiento de la matriz tridireccional que se ha considerado para la técnica PCR. Dado que la regresión lineal múltiple sólo puede realizarse con una variable respuesta, y en este caso se dispone del IDH para 14 años, es necesario "condensar" la información de estos 14 años en una sola variable. En este punto se plantean dos alternativas: o bien obtener el IDH promedio de los 14 años, o bien obtener la variable latente de dichos años, asociada a la primera componente principal. Aunque obtener la media resulta más sencillo, la segunda opción parece en este caso más recomendable.

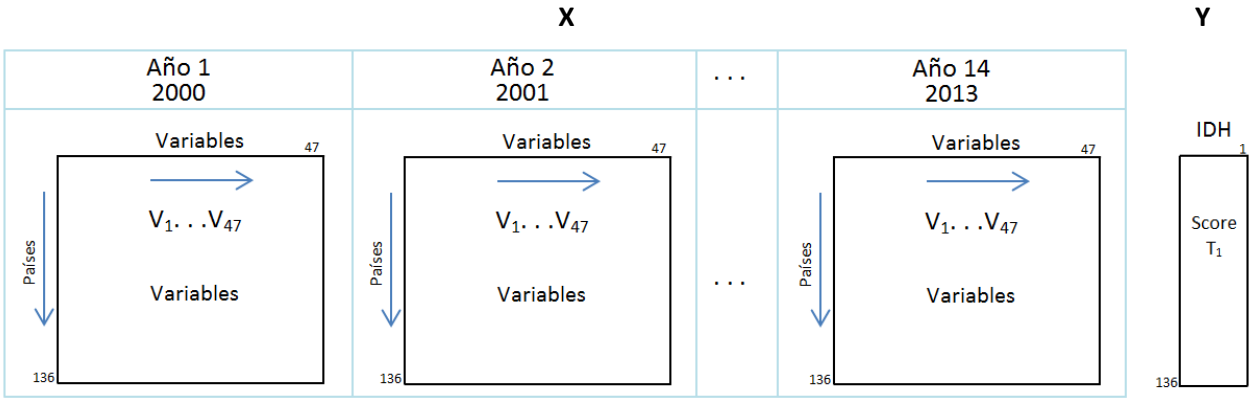

Figura 24. Despliegue bidireccional de la matriz X: 136 países por 658 variables (14 años x 47 indicadores) y *scores* T1 de IDH

Así pues, en primer lugar se ha realizado un PCA con la matriz de IDH formada por 136 observaciones y 14 valores de IDH, uno para cada país (se ha eliminado China, USA e India). Se obtiene que la primera componente principal explica el 93,0% de la variabilidad ($Q^2 = 0{,}919$), lo

cual implica que existe una alta correlación entre los valores de IDH correspondientes a los distintos años, como era de esperar.

A continuación se ha obtenido la variable latente correspondiente a la primera componente principal de esta matriz (vector de *scores* T_1 en la figura 24), lo cual básicamente sería equivalente a haber calculado el valor promedio de IDH para los 14 años.

Posteriormente, se ha realizado un PCA con la matriz X desdoblada (Figura 24) formada por 136 observaciones (sin los tres países atípicos: USA, China e India), con 47 x 14 variables. La tabla de resultados de este PCA es la siguiente:

Tabla 7. Características de las 16 componentes principales de la matriz X de la figura 24

Componentes	R^2_X	R^2_X (acum.)	Valor propio	Q^2	$Q^2_{límite}$	Q^2 (acum.)
1	0,393	0,393	53,5	0,378	0,00883	0,378
2	0,0799	0,473	10,9	0,0976	0,00888	0,438
3	0,0567	0,53	7,7	0,0642	0,00894	0,475
4	0,0493	0,579	6,71	0,0771	0,009	0,515
5	0,0384	0,618	5,22	0,0487	0,00906	0,539
6	0,0331	0,651	4,51	0,0441	0,00912	0,559
7	0,0271	0,678	3,68	0,0422	0,00918	0,578
8	0,0229	0,701	3,11	0,0307	0,00924	0,591
9	0,0204	0,721	2,78	0,0208	0,0093	0,599
10	0,0187	0,74	2,54	0,013	0,00937	0,604
11	0,0172	0,757	2,34	0,0171	0,00943	0,611
12	0,0156	0,773	2,12	0,0129	0,0095	0,616
13	0,0148	0,787	2,02	0,0143	0,00957	0,622
14	0,0136	0,801	1,85	-0,00691	0,00963	0,621
15	0,0124	0,813	1,69	-0,00212	0,0097	0,621
16	0,0117	0,825	1,59	0,0118	0,00977	0,622

Se obtiene un total de 13 componentes con Q^2 positivo, las cuales explican un 78,7% de la variabilidad de los datos. A continuación, se han obtenido con el programa SIMCA-P los *scores* correspondientes a estas 13 componentes. Todos estos valores, junto a los *scores* de la primera componente principal de la matriz IDH, se han llevado al programa Statgraphics para realizar una regresión lineal múltiple. En dicha regresión también se han incluido los cuadráticos de estas variables latentes como se muestra en la siguiente figura.

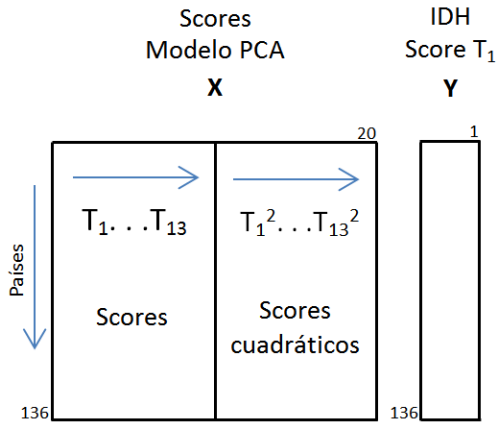

Figura 25. Estructura del modelo PCR siendo la variable respuesta el IDH (scores de la primera componente de IDH)

Los resultados de la regresión paso a paso (stepwise) son los siguientes. Se obtiene el mismo resultado con la opción *forward* y *backward*. Representando los residuos en un papel probabilístico normal se ha comprobado que no hay observaciones anómalas que requieran ser descartadas.

Multiple Regression Analysis

Dependent variable: IDH_T_1

Parameter	Estimate	Standard Error	T Statistic	P-Value
Constant	-0,379038	0,104424	-3,62979	0,0004
T_1	0,235214	0,00518828	45,3357	0,0000
T_1^2	0,00211406	0,000320137	6,60359	0,0000
T_5	0,0715218	0,0129098	5,54013	0,0000
T_3	-0,0499041	0,0114864	-4,34464	0,0000
T_8	-0,0586324	0,0163361	-3,58913	0,0005
T_2	-0,0251326	0,00897945	-2,7989	0,0059

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1675,89	6	279,315	441,09	0,0000
Residual	81,6882	129	0,633242		
Total (Corr.)	1757,58	135			

R-squared = 95,3522 percent
R-squared (adjusted for d.f.) = 95,136 percent
Standard Error of Est. = 0,795765

La siguiente figura muestra los valores de la variable latente IDH (eje "y") en función de la primera componente principal de la matriz X, es decir en función de T_1 (eje "x"). Se observa claramente un efecto cuadrático, que es recogido por la ecuación anterior ya que el p-valor del término cuadrático T_1^2 es claramente significativo ($p < 0.0001$). De hecho, estas dos variables (T_1 y T_1^2) son las dos más importantes del modelo anterior.

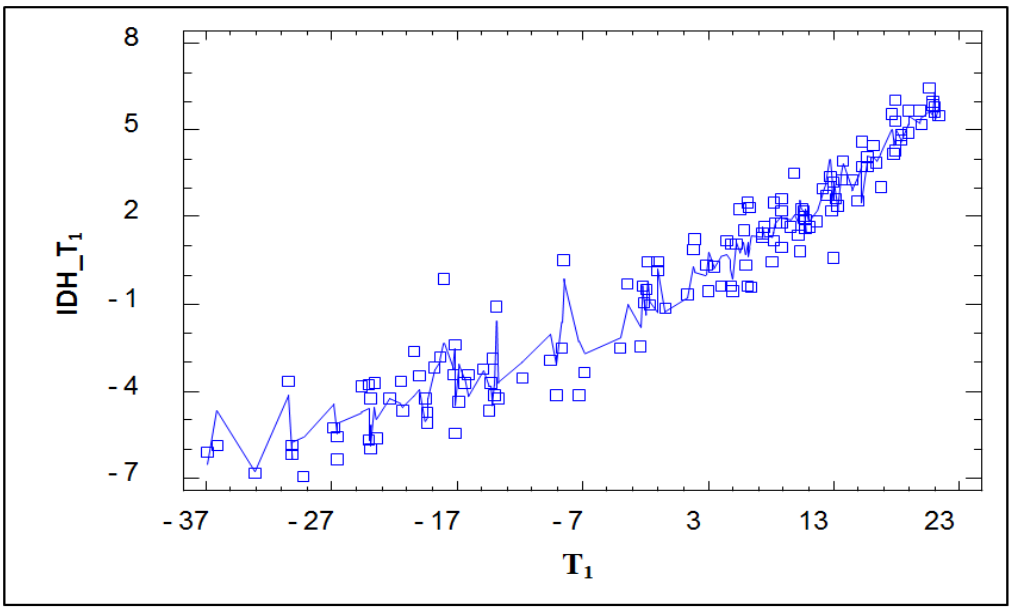

Figura 26. Gráfico de dispersión de IDH en función de T_1 (variable latente asociada a la primera componente principal de X)
Se muestran también los valores predichos (unidos con líneas)

De esta figura se deduce que no hay homocedasticidad, pues la dispersión de los residuos es mayor para valores bajos de T_1 (países menos desarrollados) que para valores altos. Si se construye un modelo solamente con las variables T_1 y T_1^2, la bondad de ajuste es relativamente buena (R^2 = 0,928). Pero este valor aumenta ligeramente hasta 0,953 al añadir adicionalmente las variables latentes T_2, T_3, T_5, y T_8, las cuales aportarán información básicamente respecto a los países menos desarrollados, pues los de mayor desarrollo se ajustan muy bien al modelo que sólo tiene T_1 y T_1^2.

A continuación se han estudiado los pesos de las variables en la formación de las componentes principales (PC) segunda, tercera, quinta y octava. El gráfico inferior muestra los pesos de PC3 frente a PC2. A partir de este gráfico se ha identificado que las variables de mayor peso (en valor absoluto) en PC3 son: "muerte por sida", "sustancias que agotan el ozono", "toneladas de CO_2" y "población total". Las variables de mayor peso en PC2 son: "porcentaje de personas con VIH", "tasa de VIH", "gasto público en instituciones públicas sobre PIB" y "tasa de incidencia de tuberculosis".

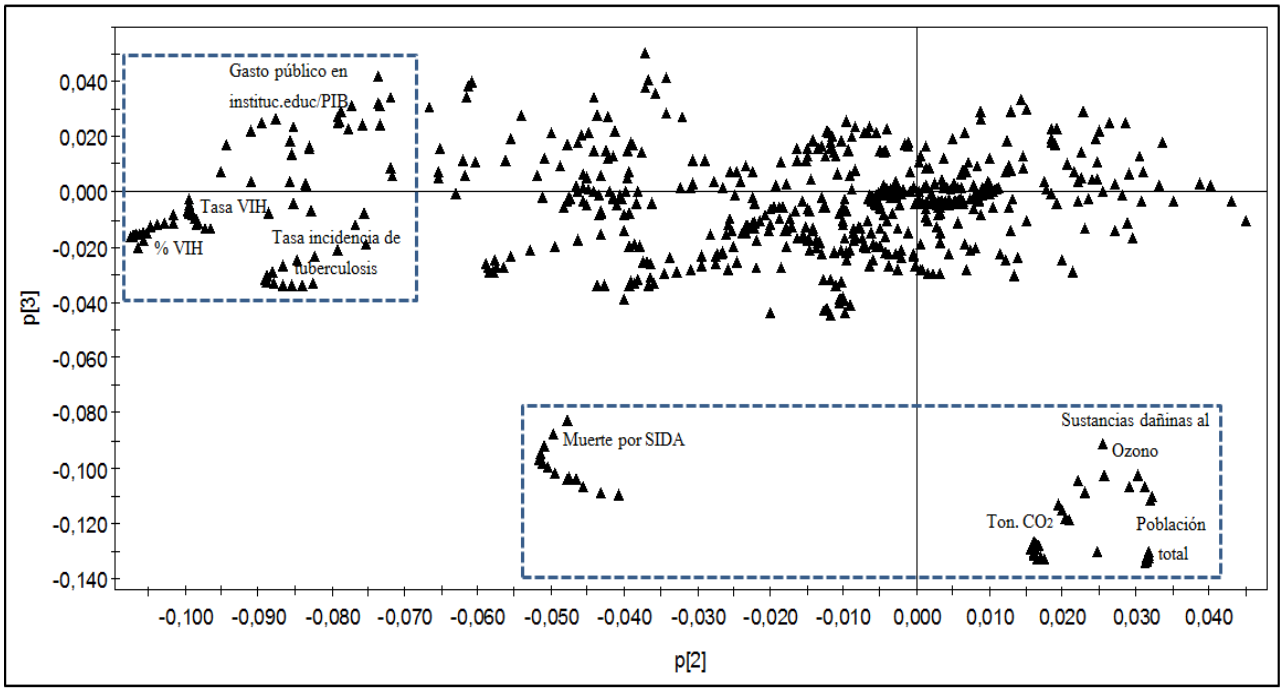

Figura 27. Gráfico de *loadings* (p_3 frente a p_2) correspondientes al modelo PCA de la matriz X de la Figura 24

El siguiente gráfico muestra los pesos de PC5 y PC8. Se observa que las variables de mayor peso (en valor absoluto) en PC5 son: "gasto en educación sobre PIB", "gasto de instituciones educativas sobre PIB", "educación sobre gasto público", "tasa de prevalencia de tuberculosis" y "mortalidad por tuberculosis". Las variables de más peso (en valor absoluto) en PC8 son: "gini", "natalidad adolescente" y "FBKF" (Formación Bruta de Capital Fijo).

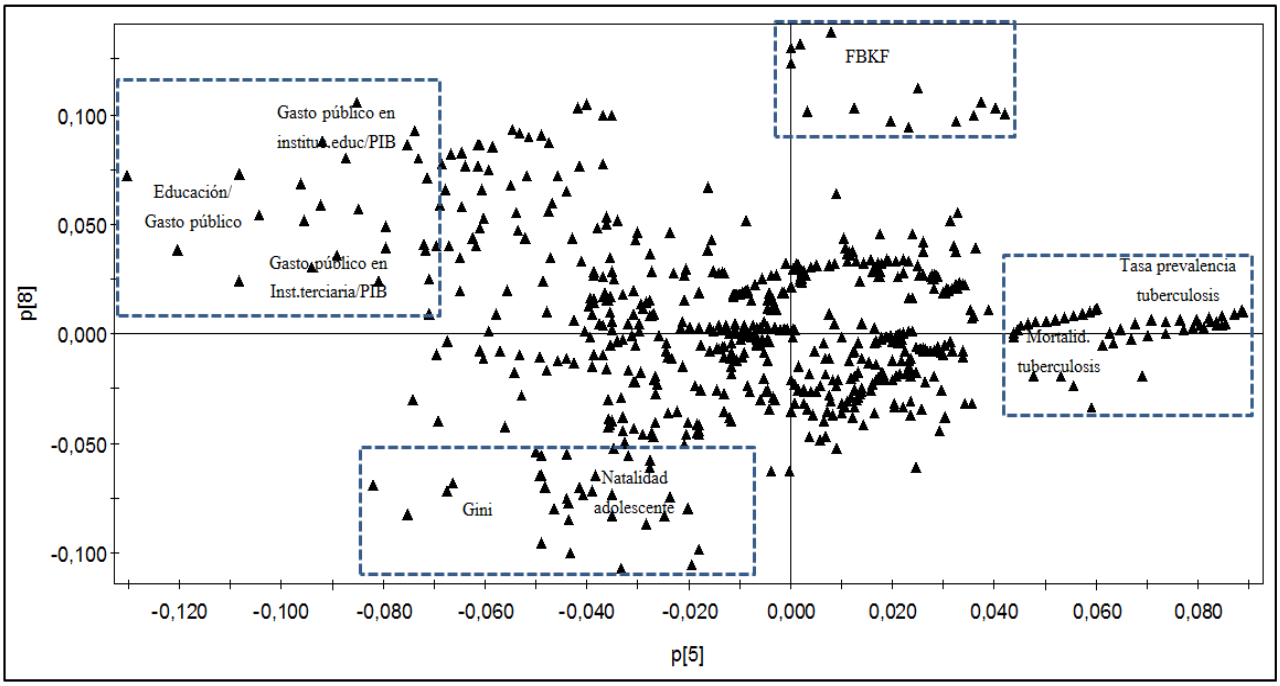

Figura 28. Gráfico de *loadings* (p_8 frente a p_5) correspondientes al modelo PCA de la matriz X de la Figura 24

A continuación se ha realizado un nuevo modelo de regresión lineal múltiple para predecir la variable latente de IDH en función de T_1 y T_1^2, de modo que el resto de variables latentes (T_2 a T_{13}) se han reemplazado por las variables de mayor peso en las componentes cuyas variables latentes aparecían en el modelo anterior y que se acaban de mencionar.

Empleando la opción de regresión paso a paso (*stepwise*), se llega a un modelo muy sencillo que es el siguiente. Este modelo elimina de forma automática 8 observaciones (países) porque los datos de alguna de las variables que entran en el modelo se desconocen.

Multiple Regression Analysis

Dependent variable: IDH_T_1

Parameter	Estimate	Standard Error	T Statistic	P-Value
Constant	-2,31077	0,449066	-5,14572	0,0000
T_1	0,273353	0,00817866	33,4227	0,0000
T_1^2	0,00320393	0,00038097	8,40993	0,0000
T_preval_tuber	0,00229272	0,000431173	5,31739	0,0000
Gini	0,0293976	0,00973468	3,01988	0,0031

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	1597,01	4	399,252	530,11	0,0000
Residual	92,6366	123	0,753143		
Total (Corr.)	1689,64	127			

R-squared = 94,5174 percent
R-squared (adjusted for d.f.) = 94,3391 percent
Standard Error of Est. = 0,867838

Este modelo tiene una bondad de ajuste prácticamente igual que el anterior (94,5% en lugar de 95,3%), pero tiene la ventaja de que su interpretación es mucho más directa. Se ha verificado que los residuos siguen una distribución normal, con una observación ligeramente anómala que corresponde al país africano Gabón. Así pues, la mayor variabilidad que se observaba en la Figura 26 para los países de menor desarrollo (que aparecen a la izquierda, con valores bajos de T_1), queda explicada en parte por las variables "gini" y "tasa de prevalencia de tuberculosis". Efectivamente, en los países menos desarrollados, la distribución de la riqueza (gini) es una variable preponderante en el índice de desarrollo humano.

3.2.2. Mínimos Cuadrados Parciales – PLS considerando Y= IDH

Para profundizar en la comprensión de las variables que más afectan el IDH, se ha empleado la regresión PLS. Con el objetivo de sacar el mejor provecho posible del modelo PLS, se han considerado dos despliegues de la matriz tridireccional X inicial. Para mayor claridad, dichos despliegues se han denominado tipo A y tipo B. El primero de ellos es una estructura bidireccional con 136 países por 658 variables (14 años x 47 indicadores). El despliegue tipo B es una matriz con 1904 observaciones (136 países x 14 años) por 94 variables. La interpretación de los resultados es distinta aunque complementaria. Así pues, a continuación se exponen los resultados obtenidos con cada uno de los desdoblamientos.

3.2.2.1. PLS, Y = IDH. Despliegue A: 136 países x 658 variables (14 años * 47 indic.)

El primer despliegue mantiene inalterada la dirección de los 136 países (observaciones) y despliega las variables a lo largo del tiempo de la forma que se muestra en la Figura 29. De esta forma, se obtiene una matriz X bidireccional con 658 variables (14 años x 47 indicadores). No obstante, dado el efecto cuadrático identificado con PCR en el apartado anterior, parece razonable incluir también los cuadrados de todas las variables. Por este motivo, en lo sucesivo, se han introducido todas las variables cuadráticas. Hay que recordar que el modelo PLS no se ve alterado demasiado si se introducen en el mismo variables redundantes, así que esta estrategia de comenzar con las 47 variables junto con sus cuadrados, parece razonable.

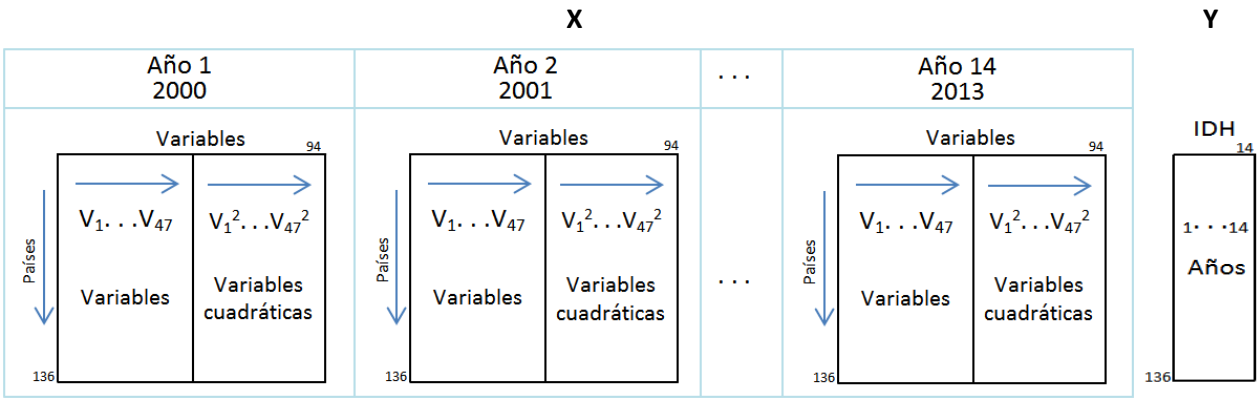

Figura 29. Despliegue tipo A de la matriz X, usado para el modelo PLS

Desdoblando la matriz tal como se indica en la figura anterior, con 136 observaciones (países), se ha realizado un PLS considerando como matriz Y las 14 variables de IDH. La tabla de resultados del modelo se muestra a continuación.

Tabla 8. Características de las diez componentes PLS correspondientes al modelo de la Figura 29

Componentes	R^2_X	R^2_X (acum.)	Valor Propio	R^2_Y	R^2_Y (acum.)	Q^2	Q^2 (acum.)
1	0,262	0,262	35,6	0,813	0,806	0,806	0,806
2	0,0647	0,327	8,8	0,0777	0,891	0,333	0,871
3	0,0561	0,383	7,63	0,0147	0,905	0,04	0,876
4	0,0402	0,423	5,47	0,00962	0,915	0,0155	0,878
5	0,0266	0,449	3,62	0,00932	0,924	0,00879	0,879
6	0,0285	0,478	3,88	0,00604	0,93	0,00394	0,879
7	0,0286	0,507	3,89	0,0039	0,934	-0,0325	0,875
8	0,0207	0,527	2,81	0,00492	0,939	-0,0164	0,873
9	0,0244	0,552	3,32	0,00323	0,942	-0,0425	0,868
10	0,018	0,57	2,44	0,00372	0,946	-0,0368	0,863

Se observa por el indicador Q^2 (positivo) que aproximadamente 6 componentes tienen capacidad predictiva del IDH. De éstas, la primera componente básicamente está explicada por los efectos simples de las variables tal como se observa en la Figura 30, y no por variables cuadráticas, lo cual tiene sentido. Las variables en la parte superior tendrán correlación positiva con IDH. Entre ellas están: "acceso a fuentes de agua mejorada", "acceso sanitario mejorado", "esperanza de vida escolar", "tasa de inicio y fin de educación primaria", "toneladas de CO_2 per cápita", "tasa de matrícula primaria" y "acceso a teléfono y celular". En la parte inferior del gráfico aparece la mortalidad infantil (niños de 0 a 1 año y de 1 a 5 años) como variables con correlación negativa.

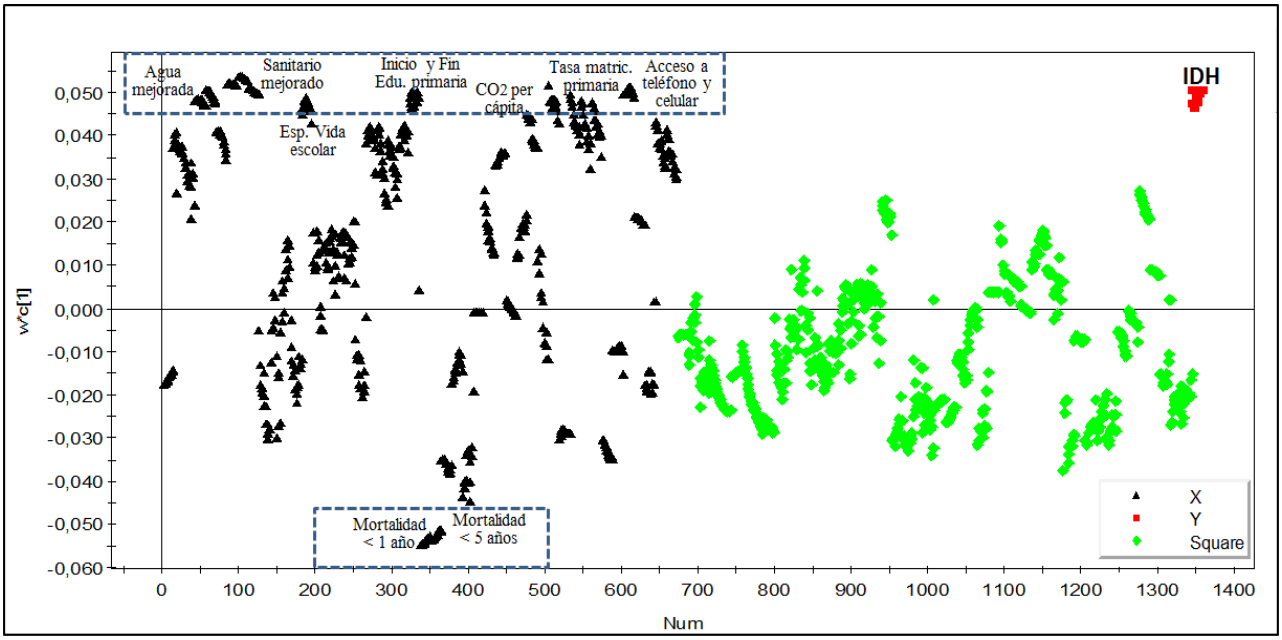

Figura 30. Loadings ($w*c_1$) de la primera componente PLS para las 1316 variables X del modelo

Respecto a la segunda componente PLS mostrada en la Figura 31, las variables no cuadráticas de mayor peso son "acceso a internet", "CO_2 per cápita" y "acceso a teléfono y celular". Aparecen algunas variables cuadráticas con un peso similar a otras variables no cuadráticas, que son [esperanza de vida]^2 y [acceso a teléfono y celular]^2.

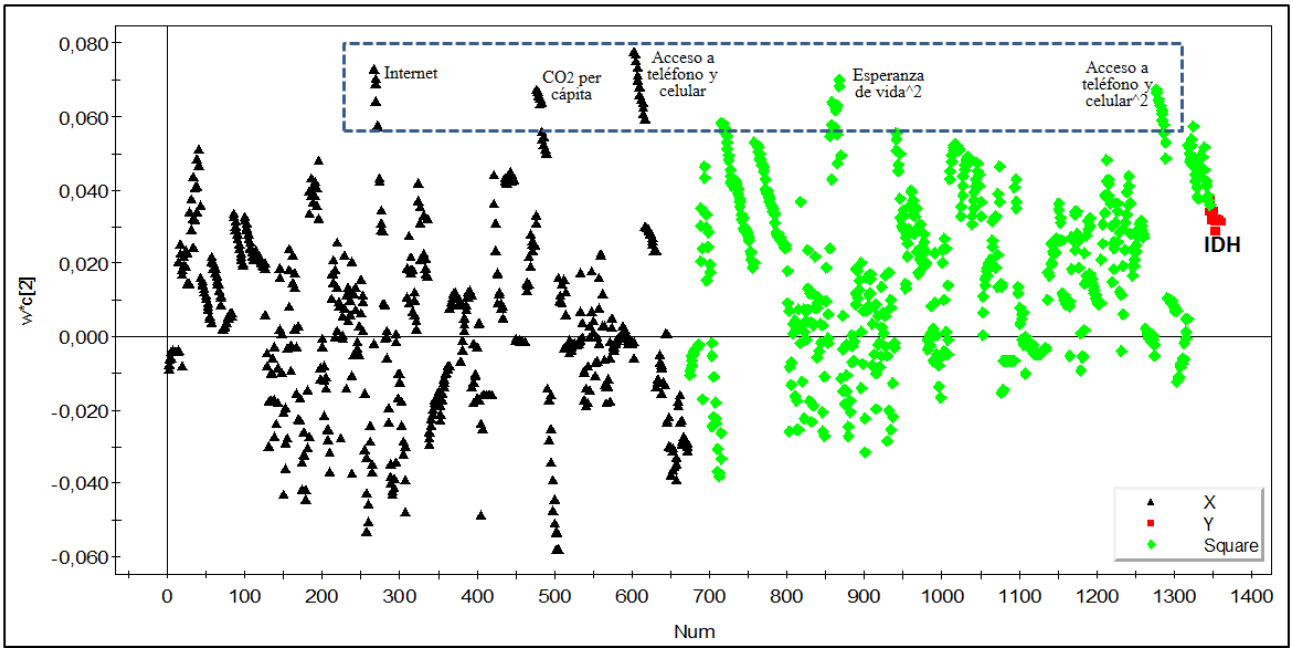

Figura 31. Loadings ($w*c_2$) de la segunda componente PLS para las 1316 variables X del modelo

La tercera componente PLS está básicamente determinada con correlación negativa por la variable "uso de energía (Kg equivalente de petróleo)". También está explicada por variables puntuales como "gini" y "gasto público en educación" de un solo año en específico. Además la variable cuadrática [mortalidad por tuberculosis]^2 también tiene importancia.

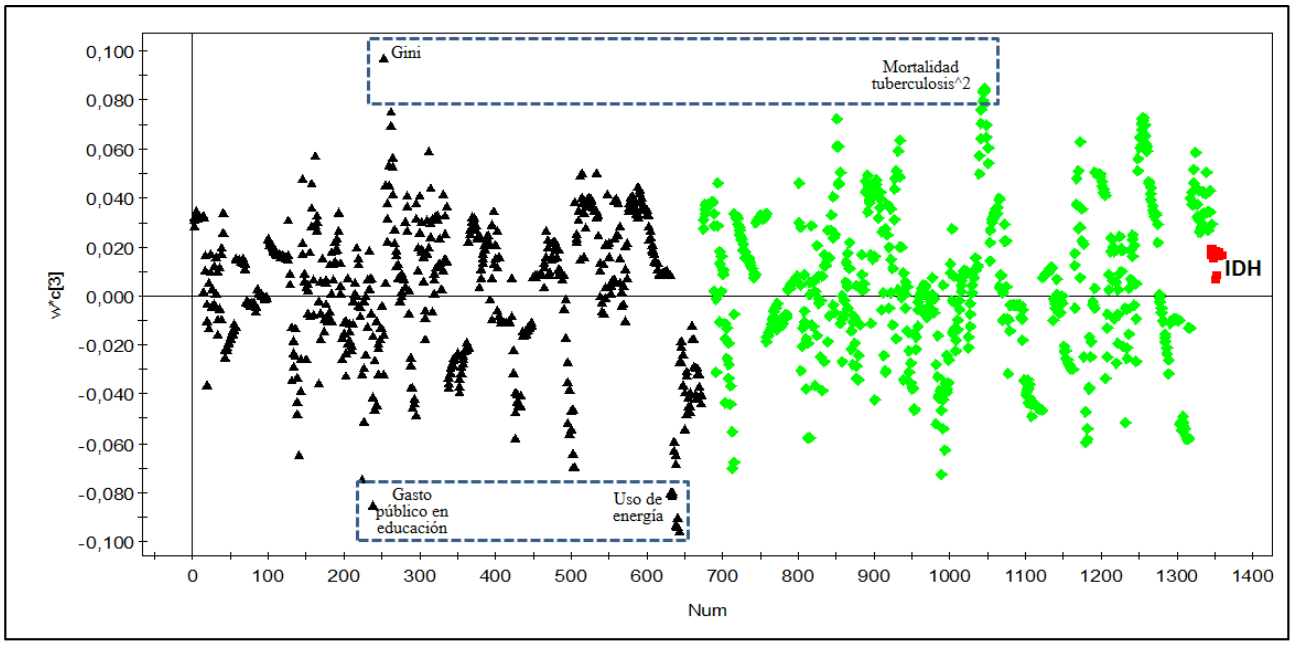

Figura 32. Loadings ($w*c_3$) de la tercera componente PLS para las 1316 variables X del modelo

En sucesivas componentes se observa que los mayores pesos son de variables "puntuales", y no de toda la secuencia de 14 años, lo cual tiene sentido. En el caso de la cuarta componente PLS, las variables no cuadráticas más importantes son "gasto público en educación sobre PIB", "gini", "crecimiento PIB", y la variable cuadrática es [gasto público en educación sobre PIB]^2.

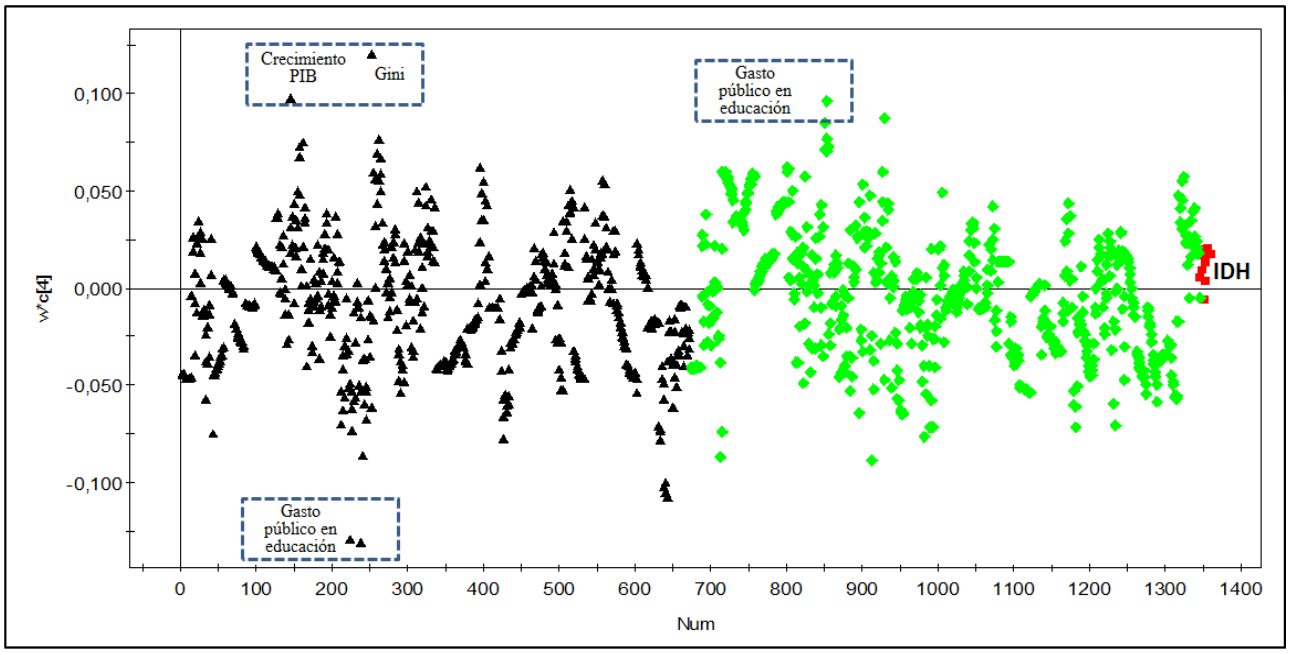

Figura 33. Loadings ($w*c_4$) de la cuarta componente PLS para las 1316 variables X del modelo

Finalmente, respecto a la quinta componente PLS, los mayores pesos corresponden únicamente a variables "puntuales" ("deuda total sobre exportaciones e importaciones de bienes y servicios" y la variable cuadrática $[gini]^2$), lo que sugiere que esta componente ya aporta muy poco a la bondad de predicción del IDH.

En resumen, básicamente las dos primeras componentes PLS explican la mayor cantidad de varianza del IDH (Tabla 8), lo cual tiene mucho sentido: la primera componente PLS puede considerarse análoga a la primera componente PCA de la matriz X, y la segunda componente PLS en cierta medida explica la curvatura del modelo. Así pues, un PLS con dos componentes tendría una bondad de ajuste R^2_Y= 0,891 (Tabla 8), la cual resulta levemente inferior al modelo PCR obtenido en el apartado anterior en la predicción de la variable latente de IDH en función de T_1 y T_1^2, cuya bondad de ajuste era R^2 = 0,928.

Las primeras 6 componentes PLS tienen un Q^2 positivo, lo que implica una bondad de ajuste acumulada R^2_Y= 0,93 (Tabla 8), la cual es levemente inferior al modelo PCR con cuatro variables (R^2= 0,945). Así pues, se obtiene que tanto la regresión PLS como PCR tienen una bondad de ajuste similar.

Al considerar un modelo PLS con 6 componentes, la Figura 34 (SPE vs. Hotelling T^2) muestra que hay tres países con valores anómalos de la T^2 considerando un nivel de confianza del 99%: África Central (CEN), Angola (ANG) y Mongolia (MON). Respecto a la distancia al modelo SPE (eje vertical), no hay ningún país anómalo considerando un nivel de confianza del 99%, pero sí 6 países ligeramente anómalos si dicho nivel es del 95%.

Cabe mencionar que Gabón (GAB en la figura 34) es el país más anómalo con arreglo a los dos criterios (SPE y T^2). Esta anomalía también se había detectado en el modelo PCR, ya que dicho país aparecía con un residuo levemente anómalo en el modelo de regresión múltiple: IDH_T_1 = f (PCA_T_1, PCA_ T_1^2, tasa_preval_tuberc, gini), mostrado en la sección 3.2.1. (pág. 42).

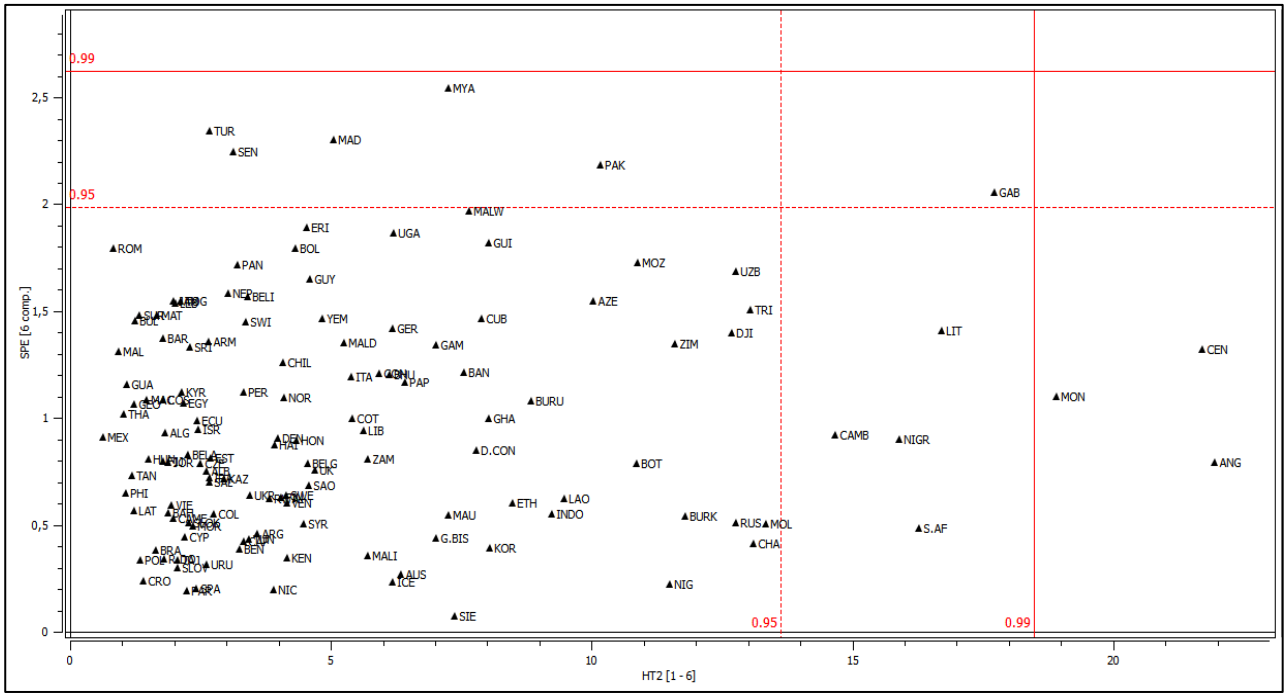

Figura 34. SPE en el espacio de las Y vs. Hotelling T^2 (modelo con 6 componentes)

El gráfico de los *scores* permite observar el posicionamiento de los países en relación al desarrollo, como se muestra en la siguiente figura. Los países de Europa muestran un mayor desarrollo humano, seguido en conjunto por los países conformados por América Latina - Caribe, Medio Oriente - Norte de África y Sur de Asia. Se puede observar que los países de África Subsahariana muestran menores niveles de desarrollo humano.

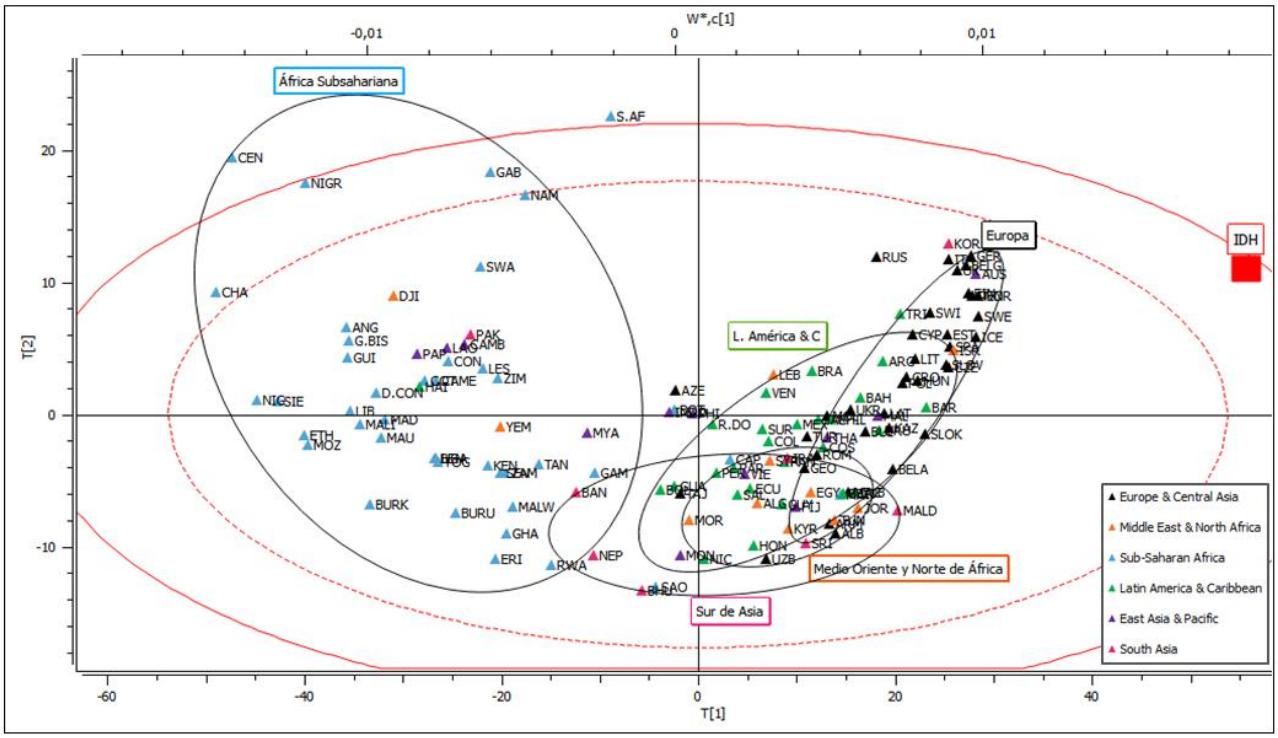

Figura 35. Gráfico biplot de los *scores* (T_2 frente a T_1) del modelo PLS. Se ha superpuesto el gráfico de pesos del IDH (C_2 frente a C_1)

El posicionamiento de los países permite identificar a la República Centroafricana (CEN) y Sudáfrica (S.AF) como países que se encuentran fuera de la T^2 de Hotelling, los cuales presentan valores bajos en "vacunas contra sarampión" y en indicadores asociados a calidad de vida como "fuente de agua mejorada" y "servicios sanitarios mejorados". Por otra parte, presentan también valores bajos en indicadores de educación e igualdad de género en educación como "porcentaje de inicio y fin de educación primaria", "esperanza de vida escolar" e "índice de paridad de género de matrícula primaria, secundaria y terciaria".

La Figura 35 muestra también que existe una relación interna no lineal, tal como se ha identificado en el modelo PCR, razón por la que se justifica la introducción de variables cuadráticas en el modelo PLS.

Al observar los *scores* en el espacio de las Y, presentadas en la siguiente figura, aparecen también agrupaciones de los países en función del grado de desarrollo. Se observa una relación similar de las regiones respecto al IDH. Algunos países se encuentran lejos de sus regiones, como es el caso de Australia, Corea del Sur (Sur de Asia), Islas Mauricio (África) e Israel (Medio Oriente). Estos países se posicionan cercanos a los países de Europa. Mientras que países como Haití (América Latina) y Pakistán (Sur de Asia) aparecen posicionados junto a países de África Subsahariana.

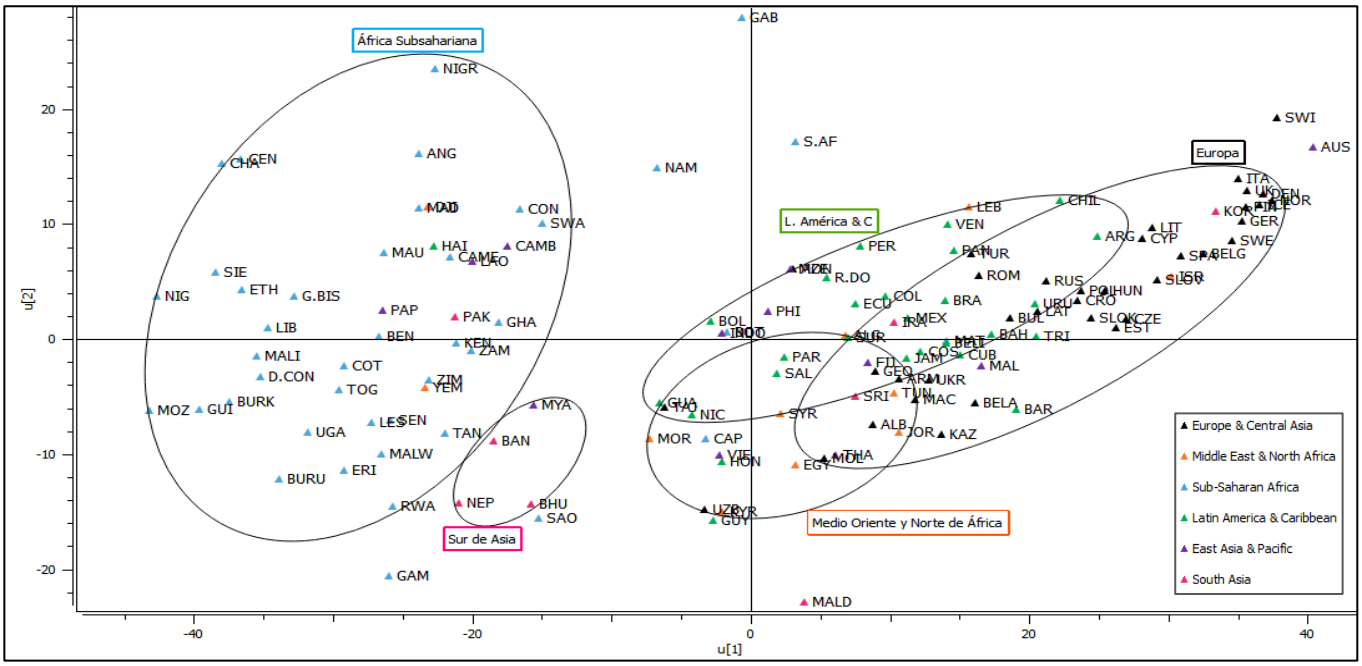

Figura 36. Gráfico de los scores en el espacio de las Y del modelo PLS del IDH (u_2 frente a u_1)

La relación existente entre el IDH y el conjunto de variables (cuadráticas o no cuadráticas) se muestra en el gráfico de los *loadings* (Figura 37). Sin embargo, ya que el despliegue considerado en este apartado permite obtener 1.316 *loadings*, uno por cada variable (94) y por cada año (14), la interpretación de este gráfico resulta algo compleja como se puede observar.

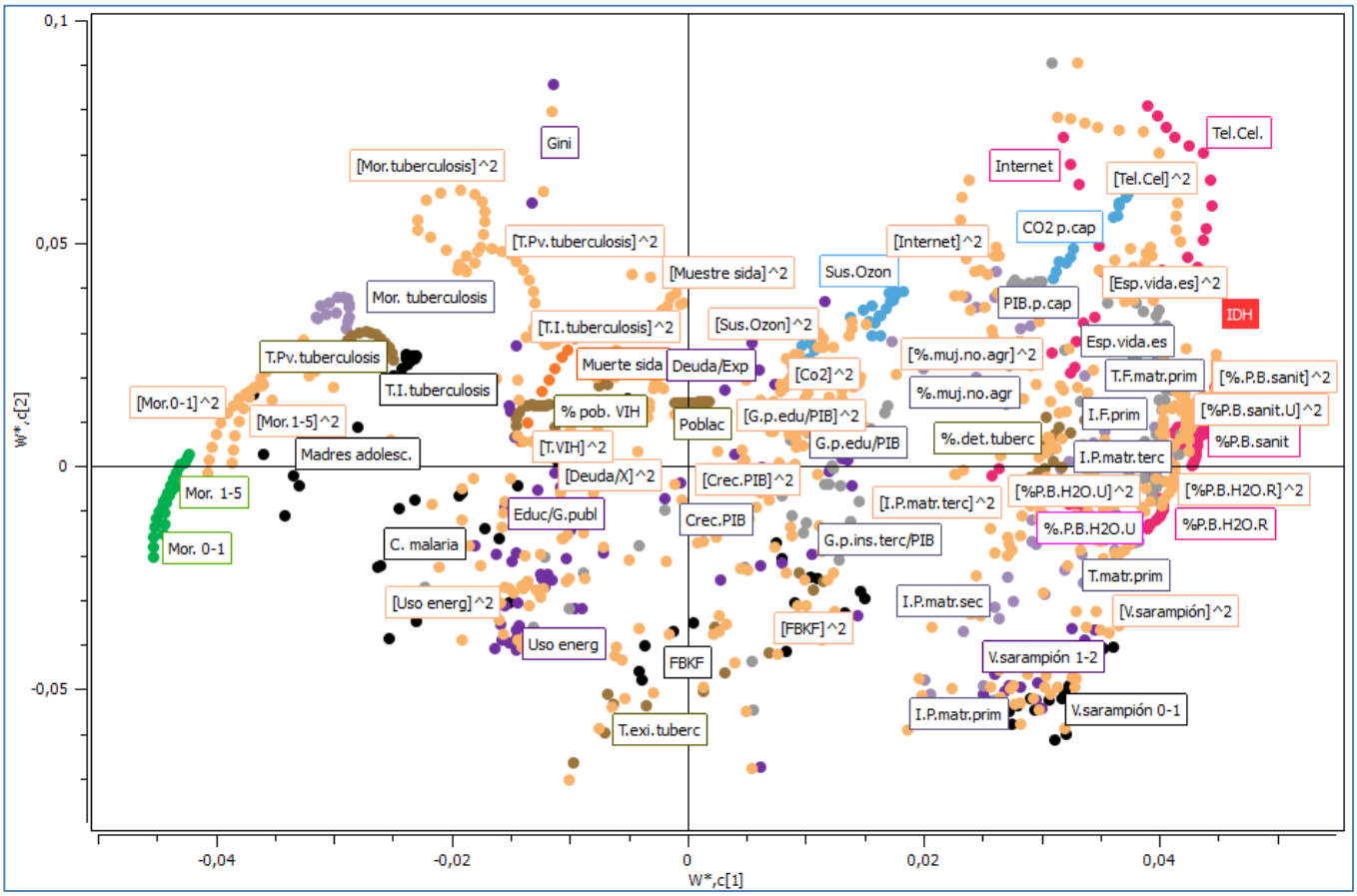

Figura 37. Gráfico de *loadings* (w*,c[2] frente a w*,c[1]) del modelo PLS con Y = IDH, con el despliegue tipo A

El gráfico de *loadings* de las variables correspondientes a la segunda y primera componentes PLS se muestra en la figura anterior. En rojo aparecen los *loadings* de las 14 variables IDH. En color marrón están las variables cuadráticas. Muchas de éstas aparecen cerca del centro, lo cual tiene sentido ya que su importancia es menor en promedio que las variables no cuadráticas. Sin embargo, hay una nube de puntos de variables cuadráticas que tienen un peso relativamente alto en la segunda componente, de magnitud similar a otras variables no cuadráticas. Esto es coherente con el hecho de que el efecto cuadrático es relevante, tal como se ha estudiado con PCR.

El gráfico muestra que el IDH mantiene una relación directamente proporcional con variables como “esperanza de vida escolar”, “tasa de inicio y fin de educación primaria”, “acceso a internet”, “porcentaje de población con acceso a agua mejorada” y “porcentaje de detección de tuberculosis”. También se observa una relación directamente proporcional con variables cuadráticas, entre ellas: “personas con acceso a agua mejorada total y rural”, “tasa de inicio y fin de educación primaria” y “tasa de matrícula primaria de niñas”.

Las variables en color morado asociadas a la igualdad de género (índice de paridad en matrícula primaria, secundaria y terciaria) que se encuentran relacionadas con el IDH, son: “esperanza de vida escolar”, “tasa de matrícula primaria para ambos sexos” y con “porcentaje de población que cuenta con fuente de agua mejorada”.

Al igual que en el modelo PCA de la sección 3.1.3, se observa una relación directa entre natalidad adolescente y mortalidad infantil. También se encuentra como es de esperar una relación directa entre "mortalidad por tuberculosis" con la tasa de prevalencia e incidencia de la tuberculosis. Lo mismo se observa para las variables relacionas con el VIH. Se deduce además que las enfermedades se encuentran inversamente relacionadas con el grupo de variables asociadas al IDH.

La correlación positiva o negativa del IDH con respecto al conjunto de todas las variables se puede observar a partir del gráfico de coeficientes de regresión del modelo PLS para cada variable con un intervalo de confianza al 95%. Estos coeficientes se han calculado sobre la matriz autoescalada. Así pues, coeficientes positivos corresponderán a variables correlacionadas positivamente con el IDH.

No obstante, el despliegue realizado en este apartado permite obtener 1316 coeficientes de regresión, uno para cada variable y para cada año, de modo que resulta difícil interpretar la siguiente figura. Por el contrario, en el desdoblamiento del tipo B que se describe a continuación, este gráfico es mucho más sencillo de visualizar e interpretar.

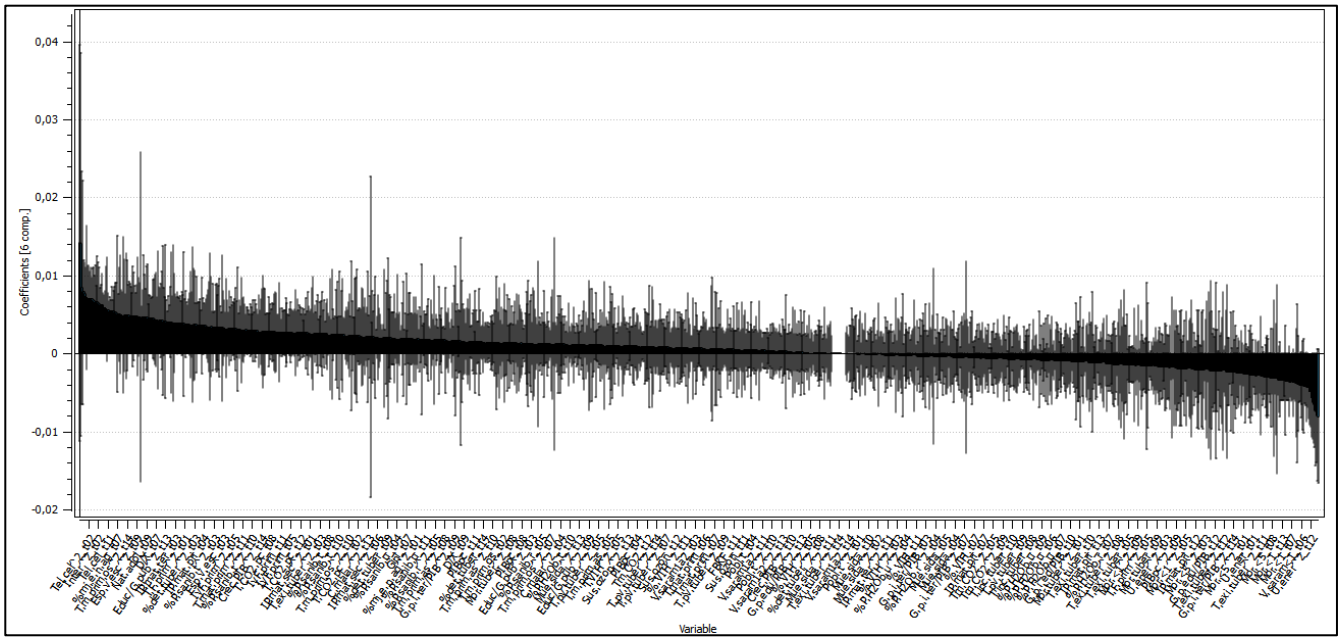

Figura 38. Coeficientes de regresión (autoescalados) del modelo PLS (Y=IDH) con despliegue tipo A (modelo con 6 componentes)

Las variables estadísticamente significativas de este modelo, es decir aquellas cuyo intervalo de confianza no contiene el valor cero, son 191, de las cuales 96 son cuadráticas. De éstas, algunas son significativas sólo para ciertos años en particular, mientras que otras variables lo son para todos los años. Entre las variables significativas con coeficientes de regresión positivos están "acceso de teléfono y celular", "acceso a internet", "esperanza de vida escolar", "tasa de inicio y fin de educación primaria" y "toneladas de CO_2 per cápita". Por el contrario, las variables más significativas con coeficientes negativos son: "mortalidad de niños menores a un año" y "mortalidad de niños entre 1 a 2 años". Estas dos variables también son las que presentan los coeficientes de regresión más negativos en el conjunto de las cuadráticas.

Entre las variables cuadráticas con coeficientes positivos cabe destacar: "porcentaje de población con acceso a fuentes de agua mejoradas", "porcentaje de personas con acceso sanitario mejorado" y "tasa de inicio y fin de educación primaria".

3.2.2.2. PLS Y=IDH despliegue B: 1904 obs. (136 países · 14 años) x 94 indicadores

El segundo despliegue mantiene inalterada la dirección de las 94 variables (incluidas las cuadráticas) y despliega los 136 países (observaciones) a lo largo del tiempo de la forma que se muestra en la Figura 39. De esta forma, se obtiene una matriz X bidireccional con 94 variables (47 cuadráticas y 47 no cuadráticas) y 1904 observaciones (136 países x 14 años).

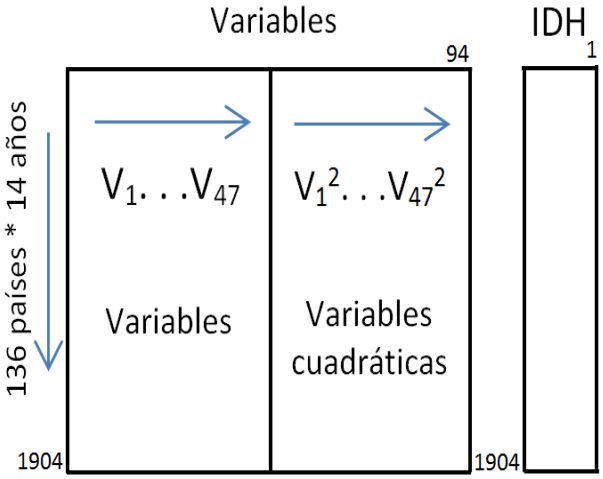

Figura 39. Despliegue tipo B de la matriz X usado para el modelo PLS

Desdoblando la matriz tal como se indica en la figura anterior, con 1904 observaciones, se ha realizado un PLS considerando el dato del IDH para cada observación. La tabla de resultados del modelo se muestra a continuación.

Tabla 9. Características de las ocho componentes del modelo PLS ajustado para el despliegue tipo B

Componentes	R^2_X	R^2_X (acum.)	Valor Propio	R^2_Y	R^2_Y (acum.)	Q^2	Q^2 (acum.)
1	0,382	0,382	35,17	0,776	0,776	0,775	0,775
2	0,050	0,432	4,61	0,048	0,824	0,194	0,819
3	0,040	0,472	3,65	0,009	0,833	0,015	0,821
4	0,044	0,516	4,01	0,005	0,838	-0,006	0,820
5	0,032	0,548	2,97	0,004	0,842	-0,012	0,818
6	0,034	0,582	3,16	0,002	0,845	-0,013	0,816
7	0,022	0,604	1,98	0,003	0,847	-0,036	0,809
8	0,028	0,631	2,53	0,002	0,850	-0,025	0,805

De este modelo PLS con despliegue tipo B, se observa en la Tabla 9 que las tres primeras componentes satisfacen el criterio de Q^2 positivo y por tanto aportan información relevante, mientras que en el modelo PLS con despliegue tipo A se observaba en la Tabla 8 que eran 6 componentes las que aportaban información relevante.

En el modelo PLS con despliegue tipo A, se alcanza una bondad de ajuste R^2_y= 0,905 con 3 componentes, mientras que en este último modelo PLS se obtiene R^2_y=0,833 considerando también 3 componentes.

Esta diferencia debida al tipo de despliegue también se observa en el indicador Q^2 que indica la bondad de predicción del modelo. En el PLS con despliegue tipo A se obtuvo acumuladamente en las tres primeras componente un Q^2= 0,876, mientras que con el despliegue tipo B se obtuvo acumuladamente en las tres componentes que aportan información relevante un Q^2 = 0,821. Estos valores no difieren demasiado, lo cual tiene sentido. De hecho, en la Tabla 8, a partir de la tercera componente, el Q^2 acumulado apenas se incrementa. Por tanto, parece concluirse que en definitiva hay tres componentes subyacentes en la predicción del IDH.

Al considerar en el modelo PLS con despliegue de datos tipo B las tres primeras componentes, la

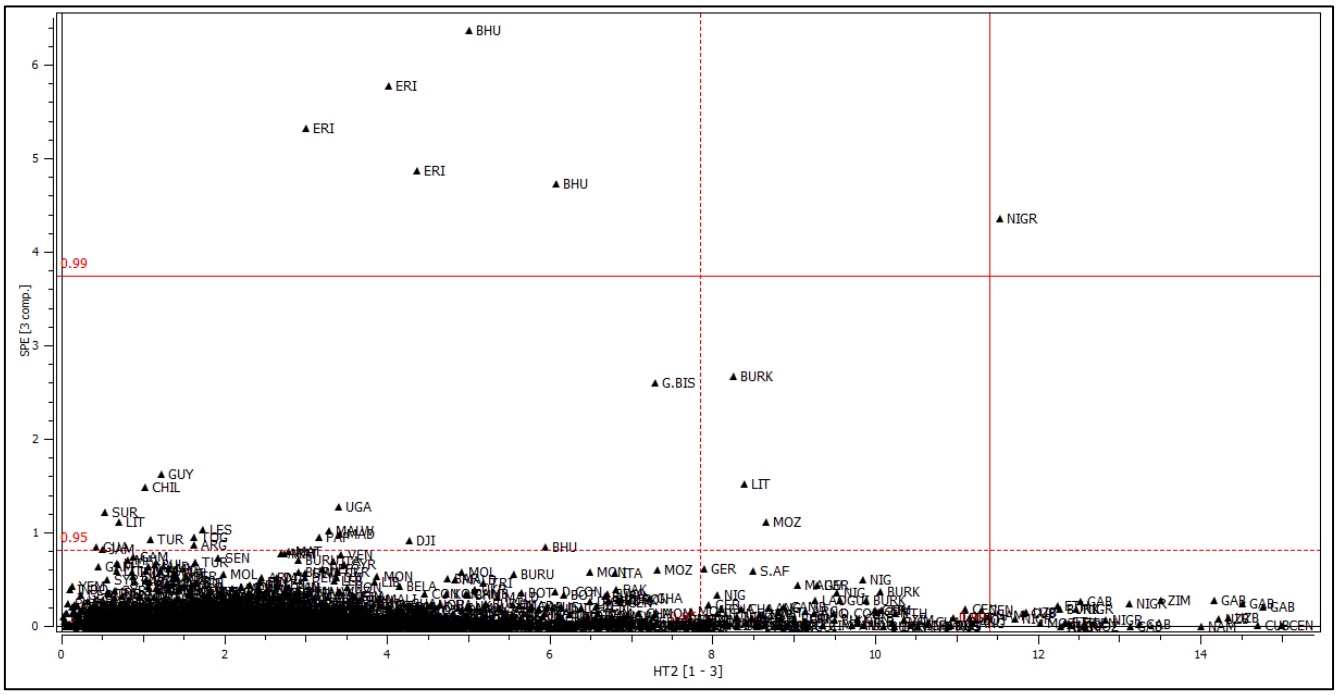

Figura 40 (SPE *vs*. Hotelling T^2) muestra que hay varias observaciones (cinco países en años específicos) con valores anómalos de la T^2 considerando un nivel de confianza del 99%: Gabón (GAB) – 6 años, África Central (CEN) –2 años, Uzbekistán (UZB) – 3 años, Mozambique (MOZ), -2 años, Nigeria (NIG) – 4 años, Cuba (CUB) – 2 años, y un solo año los países Lao (LAO), Lituania (LIT), Burkina Faso (BURK), Namibia (NAM) y Zimbabwe (ZIM). Gabón parece ser el país más anómalo en cuanto a T^2. Este hecho también se refleja en la Figura 34, como se ha comentado anteriormente.

Respecto a la distancia al modelo SPE, se observa que hay cinco observaciones (dos países en años específicos) con una leve distancia al modelo considerando un nivel de confianza del 99%.

Se observa que existe una observación que muestra anomalías en T^2 y en distancia al modelo, la cual es Nigeria (NIGR) año 2000. Esto es debido a que presenta bajos valores en porcentaje de niños menores a 2 años vacunados contra sarampión.

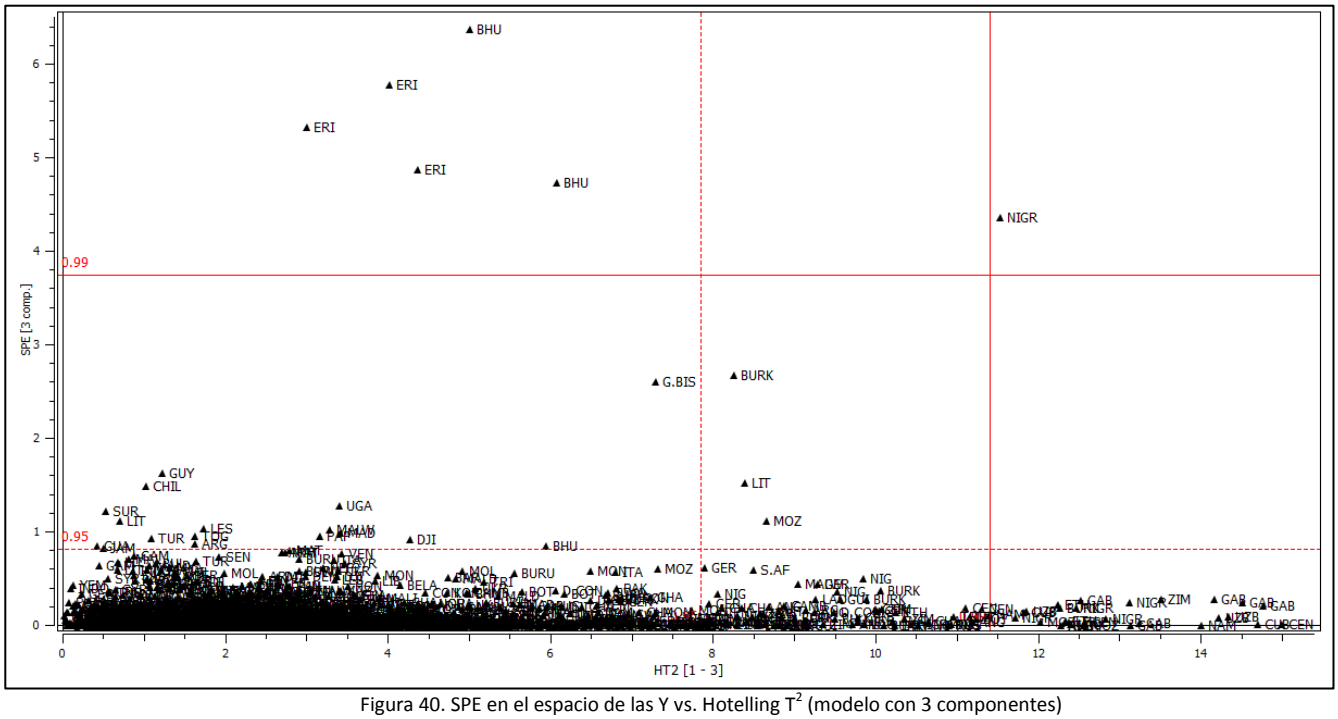

Figura 40. SPE en el espacio de las Y vs. Hotelling T^2 (modelo con 3 componentes)

A continuación resulta pertinente comparar los resultados obtenidos con PLS para cada uno de los dos despliegues propuestos. A pesar de haberse detectado algunas observaciones anómalas, éstas no se han eliminado pues realmente son pocas en comparación con las 1904 que contiene el modelo.

Lógicamente los dos tipos de despliegue dan lugar a matrices de *scores* y *loadings* de distinto tamaño. Para cada componente, en el despliegue tipo A se obtienen 136 *scores* y 658 *loadings*. Mientras que en el despliegue tipo B se obtienen 1904 *scores* y 94 *loadings*.

Esta diferencia influye en el número de observaciones que se muestran anómalas (tanto en SPE como en T^2), en la interpretación de los gráficos de *scores* y *loadings*, así como en la capacidad predictiva del modelo tal como se ha comentado.

El gráfico de *scores* del PLS con despliegue del tipo B (Figura 41) permite identificar el posicionamiento de cercanía y lejanía de los países en relación al desarrollo. Se observa cierto parecido en la posición de los países con respecto a la Figura 35 del modelo PLS con despliegue tipo A, como era de esperar. Se observa que los países de Europa muestran un mayor desarrollo humano, seguido en conjunto por los países conformados por América Latina - Caribe, Medio Oriente - Norte de África y Sur de Asia. Se puede observar que los países de África Subsahariana muestran menores niveles de desarrollo.

Entre las diferencias de los dos tipos de desdoblamiento, se puede observar que con el desdoblamiento de tipo B Nigeria (NIG) queda ubicado en la parte superior izquierda y sobresale en algunos años de la elipse T^2 de Hotelling con un nivel de confianza del 99%. Namibia también aparece para un año por fuera de esta elipse.

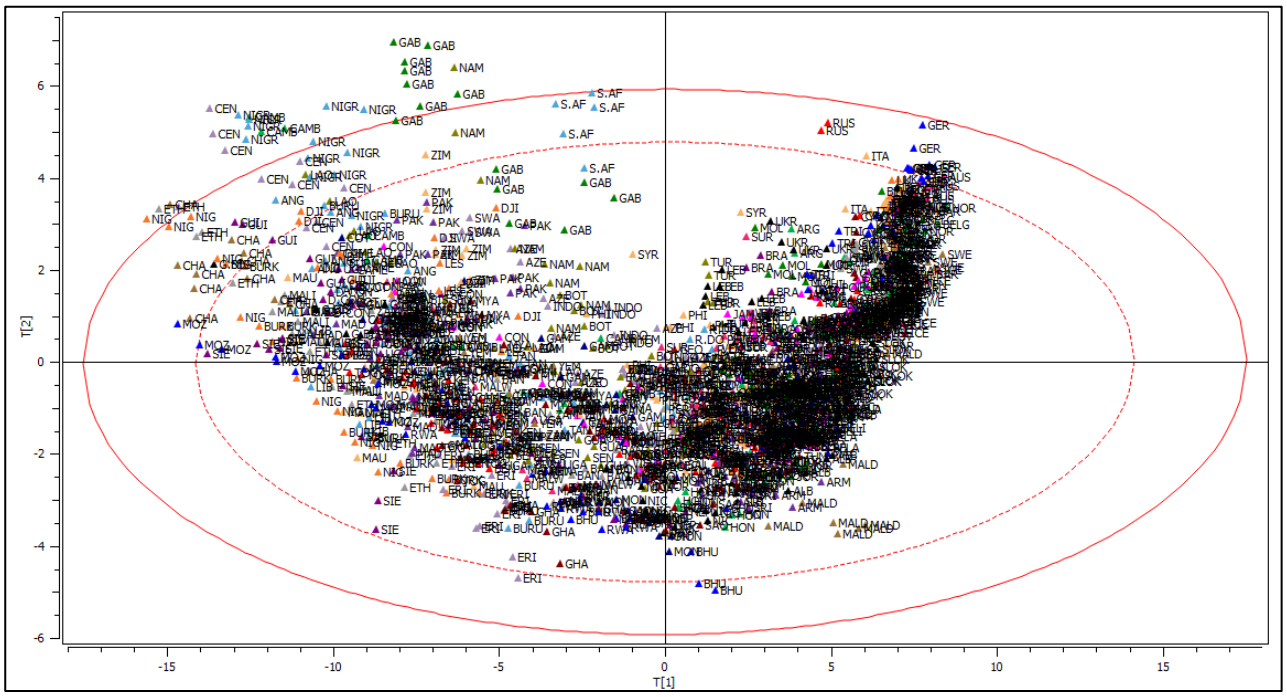

Figura 41. Gráfico biplot de los *scores* (T_2 frente a T_1) del modelo PLS. Se ha superpuesto el gráfico de pesos del IDH (C_2 frente a C_1)

Este gráfico de los *scores* resulta interesante a efectos de "control" del desarrollo humano. Cada año se publican los indicadores para la mayoría de países, los cuales serían nuevas observaciones de la matriz. Dichas observaciones podrían proyectarse sobre este modelo, y se podría evaluar su posición respecto al resto de años para ese país. En caso de que una observación cayese fuera de la elipse T^2 de Hotelling sería también una señal. Así se podría focalizar de mejor manera la ayuda a un país en específico en caso que se detecte cualquier anomalía.

Se observa en el gráfico de los *scores* (Figura 41) la misma relación interna no lineal que se ha identificado en el modelo PCR y el modelo PLS con desdoblamiento del tipo A.

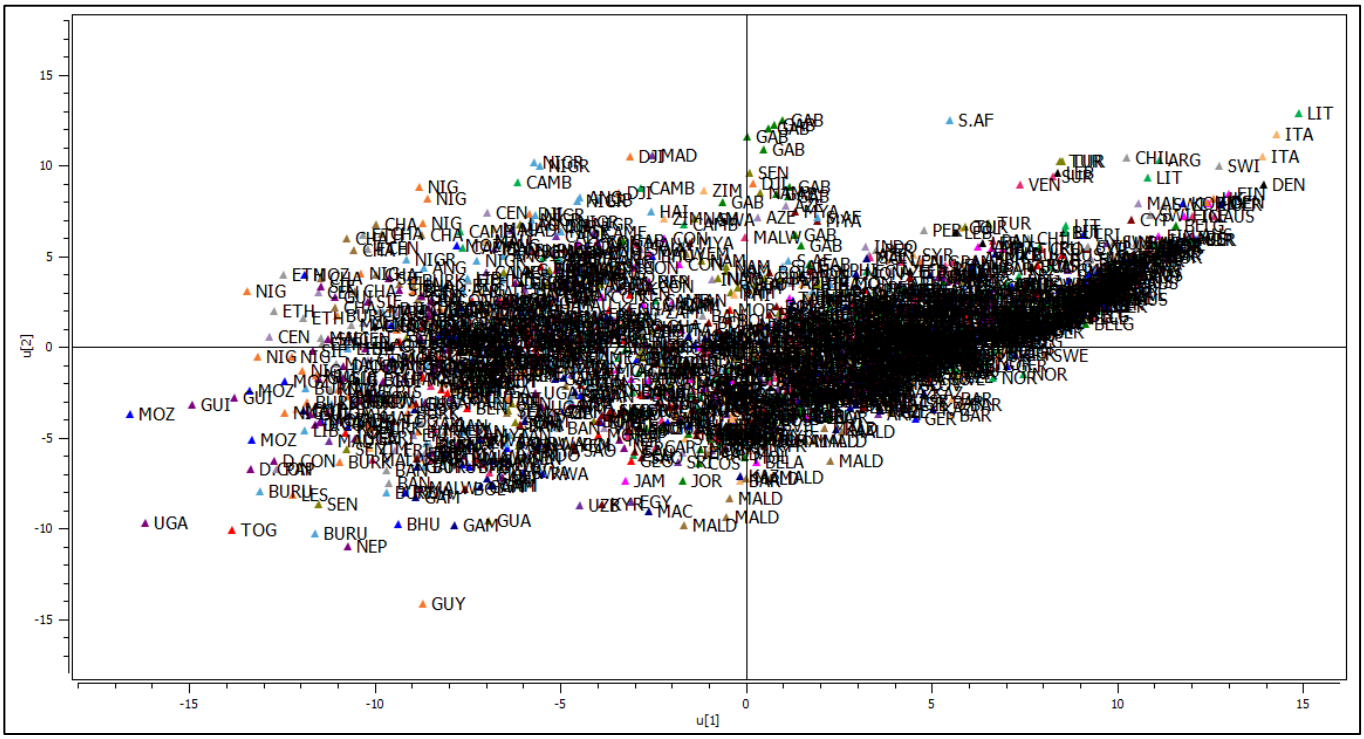

Figura 42. Gráfico de los *scores* en el espacio de las Y del modelo PLS del IDH (u_2 frente a u_1) con despliegue tipo B

Resulta difícil analizar el gráfico de los *scores* en el espacio de las Y, ya que aparecen 1904 observaciones. Este gráfico es comparable con la Figura 36 correspondiente al PLS con desdoblamiento del tipo A, en el cual se visualiza de forma mucho más clara la posición de los distintos países.

Sin embargo, una ventaja del despliegue tipo B es la interpretación del gráfico de *loadings*, ya que con el despliegue tipo A (Figura 37) aparecen 1.316 *loadings* (94 variables x 14 años), mientras que con el despliegue tipo B se observan solamente 94 *loadings*, como se puede observar en la siguiente figura. Esto facilita bastante la interpretación del gráfico.

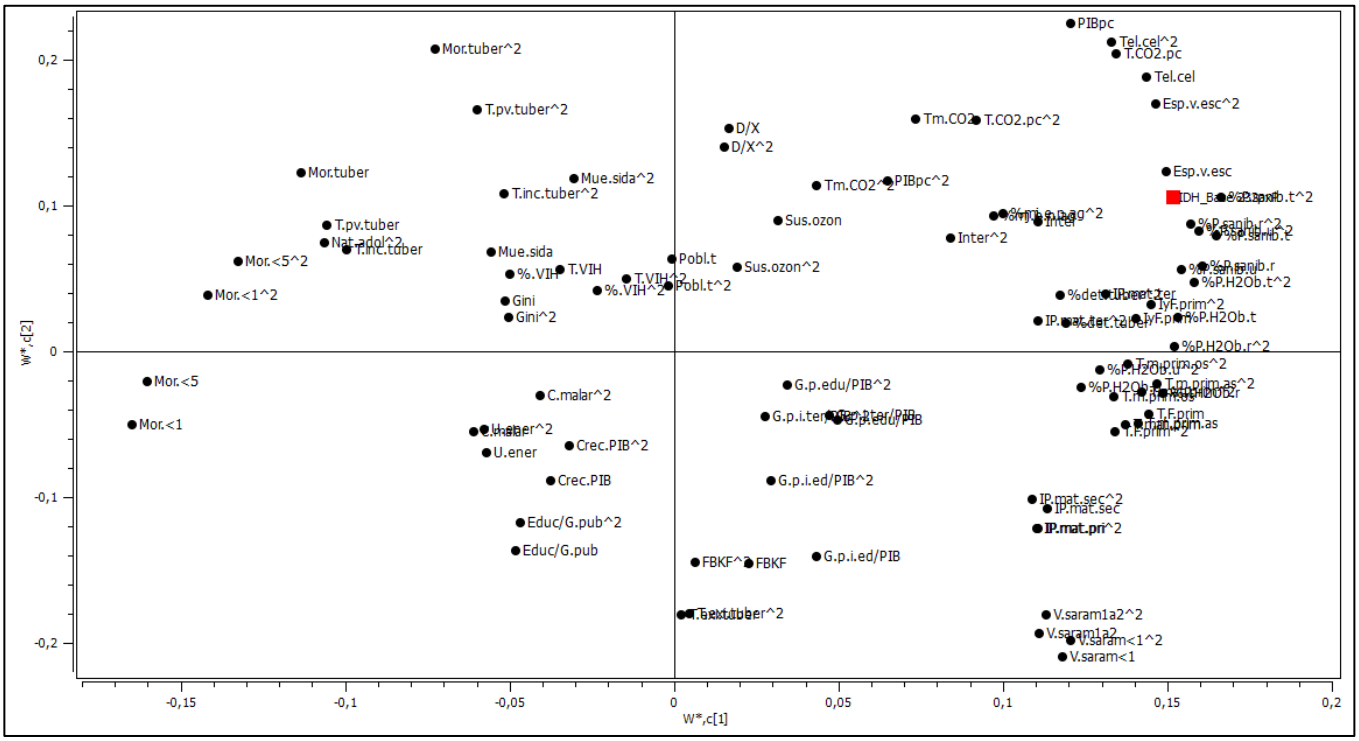

Figura 43. Gráfico de *loadings* (w*,c[2] frente a w*,c[1]) del modelo PLS con Y = IDH, con el despliegue tipo B

El gráfico de *loadings* correspondiente a la primera y segunda componentes PLS con despliegue tipo B que se muestra en la figura anterior muestra en rojo (parte superior derecha) al IDH. Este gráfico muestra relaciones muy similares de las variables que las presentadas en el gráfico de *loadings* del PLS con despliegue tipo A (Figura 37). Se observa que muchas de las variables cuadráticas aparecen cerca del centro. Esto cual implica una importancia menor respecto a las variables no cuadráticas, lo cual es bastante lógico.

Sin embargo, aparece una nube de puntos de variables cuadráticas en la parte superior derecha e izquierda, que tienen un peso relativamente alto en la segunda componente, de magnitud similar a otras variables no cuadráticas. Esto evidencia el efecto cuadrático relevante, tal como se ha estudiado anteriormente por medio de PLS con despliegue tipo A y también con PCR.

El gráfico muestra que el IDH mantiene una relación directamente proporcional con variables como “esperanza de vida escolar”, “tasa de inicio y fin de educación primaria”, “acceso a internet”, “porcentaje de población con acceso a agua mejorada” y “porcentaje de población con servicio sanitario mejorado”. Así también se observa una relación directamente proporcional con variables cuadráticas, entre ellas: “porcentaje de población con servicio sanitario mejorado”, “población con acceso a agua mejorada total y rural”, “esperanza de vida escolar” y “tasa de inicio y fin de educación primaria”.

Se muestra en el gráfico, al igual que con el PLS con despliegue tipo A, que existe una asociación de variables relacionadas con la igualdad de género (índice de paridad en matrícula primaria y secundaria). También se observa una correlación positiva entre maternidad adolescente (Nat.adol), tasa de prevalencia de tuberculosis (T.pv.tuber) y mortalidad infantil.

La correlación del IDH con respecto al conjunto de variables se puede deducir a partir del gráfico de coeficientes de regresión del modelo PLS. Estos coeficientes mostrados en la Figura 44 se han calculado sobre la matriz autoescalada con un intervalo de confianza del 95%.

El despliegue realizado en este apartado permite obtener 94 coeficientes de regresión, uno para cada variable, lo cual facilita el diagnóstico del modelo al observar las variables con un coeficiente estadísticamente significativo. Este gráfico de coeficientes tiene una gran ventaja de interpretación respecto al gráfico de coeficientes obtenido con desdoblamiento tipo A (Figura 38).

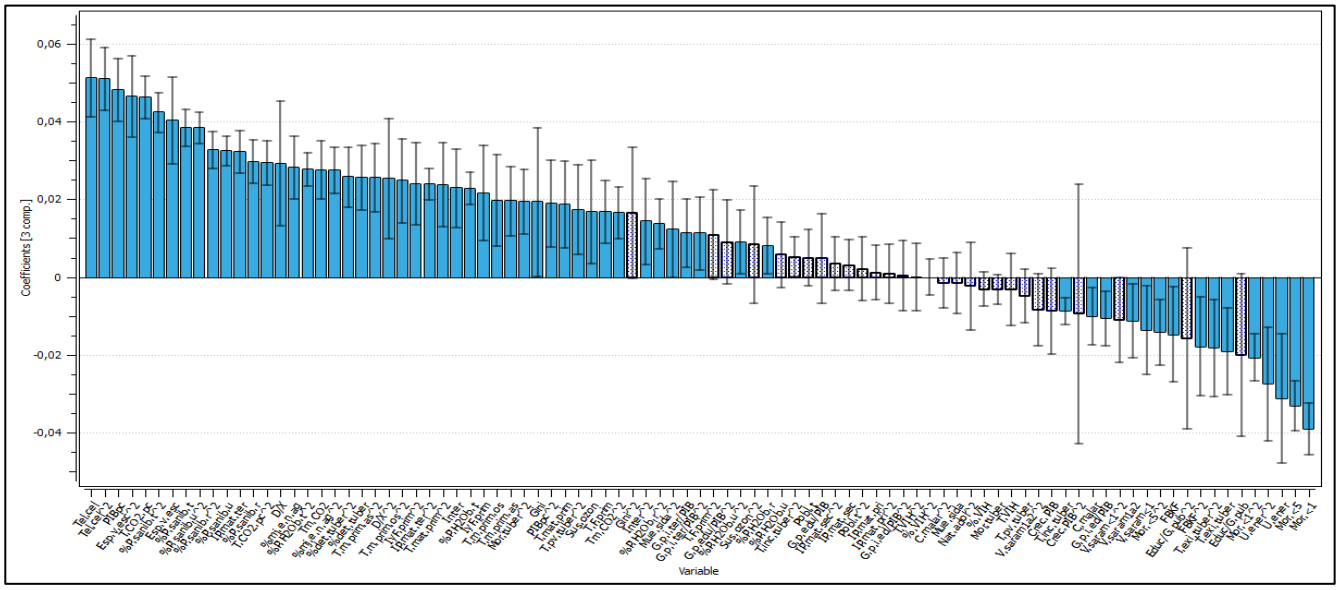

Figura 44. Coeficientes de regresión (autoescalados) del modelo PLS con despliegue tipo B, obtenido con 3 componentes

El gráfico de coeficientes de regresión de la figura anterior, obtenido para el modelo con 3 componentes, indica que las variables estadísticamente significativas son 65, pintadas de color azul, es decir aquellas cuyo intervalo de confianza no contiene el valor cero.

Entre las variables significativas con coeficientes de regresión positivos están “acceso de teléfono y celular”, “PIB per cápita”, “esperanza de vida escolar”, “toneladas de CO_2 per cápita”, “porcentaje de población con servicio sanitario mejorado”, “porcentaje de personas con acceso a fuentes mejoradas de agua”, “índice de paridad en matrícula de tercer nivel”, “deuda sobre exportaciones de bienes y servicios”, “porcentaje de mujeres empleadas en el sector no agrícola”, “acceso a internet” y “Gini”. Por el contrario, las variables más significativas con coeficientes negativos son: “mortalidad de niños menores a un año”, “mortalidad de niños entre 1 a 5 años”, “uso de energía (kg equivalentes a petróleo) por $1.000 PIB” y “FBKF”.

Entre las variables cuadráticas con coeficientes positivos cabe destacar: “acceso de teléfono y celular”, “esperanza de vida escolar”, “porcentaje de personas con acceso sanitario mejorado nacional y urbano”, “porcentaje de población con acceso a fuentes de agua mejoradas” y “porcentaje de mujeres empleadas en el sector no agrícola”. Y las variables cuadráticas con coeficientes negativos: “uso de energía (kg equivalentes a petróleo) por $1.000 PIB”, “mortalidad de niños menores a un año”, “FBKF” y “mortalidad de niños entre 1 a 5 años”.

3.3. Estimación del IDH para los países sin este valor

Los anteriores modelos predictivos de IDH basados en PCR y PLS permiten estimar este parámetro para los países que no cuentan con dicho valor. Esto ocurre con ciertos países en los cuales las Naciones Unidas no disponen de suficiente información para calcular directamente el IDH. Sin embargo, los modelos mencionados PCR y PLS son capaces de pronosticar el IDH en función de las pocas variables que se conocen de esos países.

Predicción del IDH en 2013 por medio de PCR

El modelo descrito en la sección 3.2.1 (pág. 39) permite predecir IDH_T_1 en función de las variables latentes de la matriz X. IDH_T_1 es la variable latente asociada a la primera componente principal de la matriz de IDH con los valores para los 14 años considerados. Se obtuvo que IDH_T_1 se relaciona con las variables latentes T_1, T_1^2, T_5, T_3, T_8, y T_2. El modelo predictivo obtenido es el siguiente:

$$IDH\_T_1 = -0{,}379 + 0.235 \cdot T_1 + 0{,}0021 \cdot T_1^2 + 0{,}0715 \cdot T_5 - 0{,}0499 \cdot T_3 - 0{,}0586 \cdot T_8 - 0{,}0251 \cdot T_2$$

Este modelo se obtuvo para los 136 países estudiados en este trabajo de los cuales se dispone del IDH, tras eliminar China, India y Estados Unidos. Sin embargo, la matriz tridireccional inicial consta de 234 países, y para muchos de ellos hay suficientes datos para estimar las variables latentes T_1, T_2, T_3, T_5 y T_8 de la matriz X.

Para realizar la predicción de los países que no cuentan con IDH, se deben calcular sus *scores* del modelo PCA. Luego se puede aplicar la ecuación anterior para estimar IDH_T_1, que es la primera componente de la matriz de IDH. Esta variable latente sintetiza la información de los 14 años, pues existe una alta correlación entre ellos. Dado que la variable latente IDH_T_1 resume la información de todos esos 14 años, previsiblemente la correlación con cada uno de los años será elevada. Esto se ha verificado para el año 2013 por medio de regresión lineal simple. La salida obtenida con Statgraphics se muestra a continuación:

Simple Lineal Regression – IDH_2013 vs. T_1

Parameter	Estimate	Standard Error	T Statistic	P-Value
Constant	0,666452	0,00201792	330,268	0,0000
IDH_T_1	0,0429905	0,000561321	76,5881	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	3,24832	1	3,24832	5865,74	0,0000
Residual	0,074206	134	0,000553778		
Total (Corr.)	3,32253	135			

Correlation coefficient = 0,98877
R-squared = 97,7666 percent
R-squared (adjusted for d.f.) = 97,7499 percent
Standard Error of Est. = 0,0235325

La regresión muestra que existe una alta correlación (R^2 = 0,978) entre la variable latente IDH_T_1 y el IDH del año 2013, como era de esperar. El gráfico de valores observados frente a predichos pone en evidencia la elevada correlación.

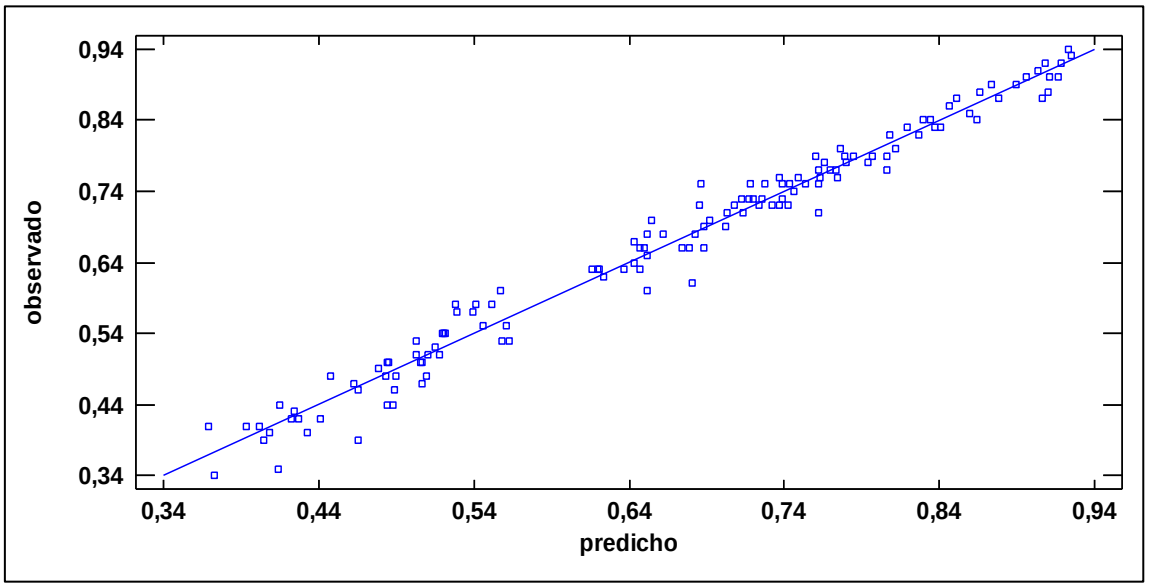

Figura 45. Valores observados frente a predichos obtenidos con regresión lineal simple: IDH_2013 = f (IDH_T_1)

Esta última regresión lineal simple permite pronosticar el IDH del año 2013 en función de la variable latente IDH_T_1, por medio de la siguiente ecuación:

$$IDH\_2013 = 0{,}666 + 0{,}04299 \cdot IDH\_T_1$$

Partiendo de la matriz tridireccional inicial con 234 países, se ha seleccionado un conjunto de 35 países de los cuales no se conoce su IDH pero se dispone de suficientes datos para poder estimar las variables latentes. El procedimiento empleado ha sido el siguiente. En primer lugar, a partir de los *loadings* de cada componente, se han obtenido por medio de una hoja de cálculo Excel los *scores* T_1, T_2, T_3, T_5 y T_8. A continuación, aplicando la ecuación de la página anterior, se ha obtenido IDH_T_1. Finalmente, a partir de este valor se ha estimado IDH_2013.

Los resultados de la predicción para estos 35 países se muestran en la siguiente tabla. Conviene recordar que los valores fluctúan en una escala de 0 a 1, por lo que no tiene sentido valores superiores a 1, como se ha obtenido para Faeroe Island y Wallis & Futuna Island. La predicción en ambos casos sería, por tanto, el valor 1.

Tabla 10. Predicción del IDH en 2013 para 35 países a partir de los modelos PCR y PLS

País	PCR	PLS	País	PCR	PLS
American Samoa	0,761	0,605	Mónaco	0,909	0,778
Anguilla	0,875	0,818	Montenegro	0,823	0,731
Aruba	0,865	0,769	Montserrat	0,804	0,774
Bermuda	0,988	0,854	Nauru	0,713	0,632
British Virgin Island	0,863	0,787	New Caledonia	0,893	0,741
Cayman Island	0,954	0,814	Niue	0,878	0,742
China, Macao	0,837	0,894	Mariana Island	0,769	0,660
Cook Island	0,848	0,695	Puerto Rico	0,858	0,693
Faeroe Island	1,311	0,869	Reunion	0,874	0,716
French Guiana	0,719	0,713	San Marino	0,838	0,749
French Polynesia	0,855	0,724	Somalia	0,393	0,534
Greenland	0,865	0,749	State of Palestine	0,683	0,652
Guadeloupe	0,839	0,703	Tokelau	0,850	0,681
Guam	0,864	0,700	Turks and Caicos Island	0,719	0,644
North Korea	0,622	0,650	Tuvalu	0,676	0,642
Malta	0,844	0,728	U.S. Virgin Island	0,877	0,730
Marshall Island	0,667	0,655	Wallis & Futuna Island	0,734	1,100
Martinique	0,831	0,666			

En general los valores obtenidos son elevados excepto para Somalia. La tabla anterior muestra también las predicciones obtenidas con PLS, según el método descrito a continuación.

Predicción del IDH en 2013 por medio de PLS

El modelo PLS permite proyectar los países que no cuentan con IDH sobre el plano del modelo y obtener una estimación de IDH_2013 para cada país. En la sección 1.4 de este trabajo se descartó un conjunto de países e indicadores por falta de información. Como resultado de ese proceso se redujo de 234 países a 139. Sin embargo, un conjunto de estos países no incluidos en los modelo del trabajo (46 en total) disponen de datos de IDH. Por ello, se considera a esos países como un conjunto de observaciones empleadas como validación externa del modelo. Así pues, para evaluar la bondad de predicción del PLS, se han proyectado estos 46 países junto con los 35 de los cuales no se dispone de IDH (indicados en la Tabla 10). En total son 81 países no empleados para ajustar el modelo PLS.

La Figura 46 muestra la distancia de estos 81 países respecto al modelo PLS ajustado con tres componentes. Se observa que ningún país muestra anomalía, ya que aparecen por debajo del límite de confianza del 95%. Esto sugiere que el modelo es adecuado.

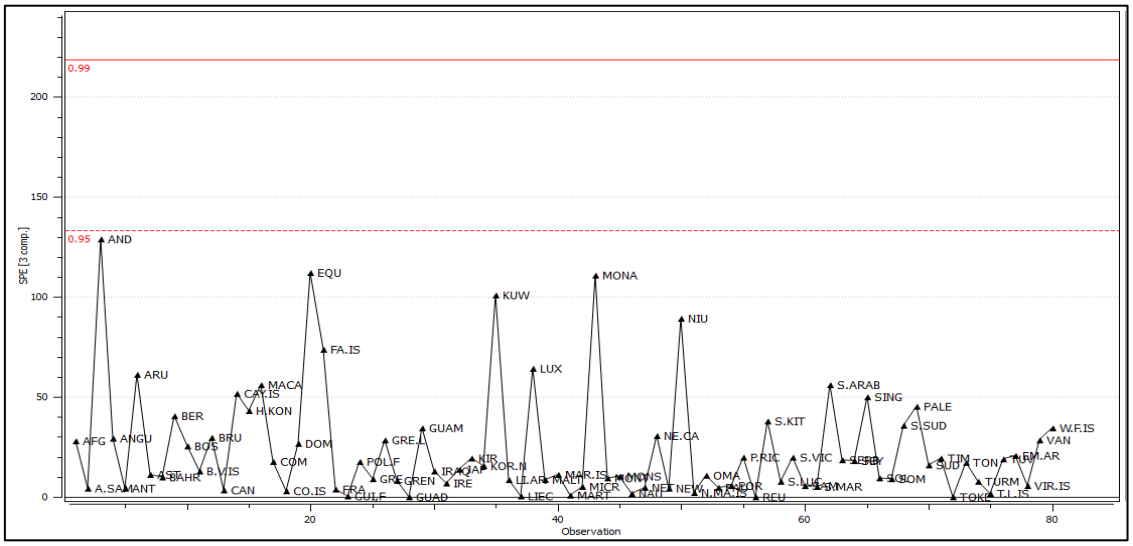

Figura 46. Distancia al modelo de los países que no cuentan con IDH para el año 2013 y países de validación externa

En el caso del modelo PLS con despliegue tipo B, permite obtener una predicción de IDH para cada año en específico. Hay que tener en cuenta que la regresión PLS se aplica después de un adecuado pretratamiento de los datos, que en este caso era centrado y escalado a varianza unitaria (datos autoescalados). Por ello, la aplicación del modelo obtenido con PLS estima los datos autoescalados de la variable IDH. Para reconstruir los valores originales de las predicciones hay que "deshacer" el pretratamiento, teniendo en cuenta que:

$$Z = \frac{X - \bar{X}}{\sigma}$$

Donde,

Z es el valor predicho de IDH_2013

$\bar{X}$ es el valor medio de IDH

σ es la desviación estándar de los valores de IDH

Los valores estimados con PLS para el parámetro IDH correspondiente al año 2013 para el conjunto de 35 países que no cuentan con esta información se muestran en la Tabla 10.

Comparación de valores IDH estimados en 2013 a partir de los modelos PCR y PLS

La Tabla 11 muestra los valores observados de IDH para el año 2013 para el conjunto de 46 países empleado como validación externa, así como las predicciones estimadas con PLS. Se han obtenido también las predicciones con el método PCR para poder comparar ambos métodos. Al contar con un valor observado y uno predicho se puede calcular un estadístico de bondad de ajuste tal como el RMSE (Root Mean Square Error), siendo Y_t el valor observado, F_t el valor predicho y *m* el número de predicciones:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{m}(Y_t - F_t)^2}{m}}$$

Tabla 11. 46 países empleados para validación externa cuyo IDH es conocido en 2013, frente al valor predicho con PCR y PLS

País	IDH 2013	Predicción PCR	Predicción PLS
Afganistán	0,464	0,436	0,551
Andorra	0,844	0,881	0,745
Antigua y Barbuda	0,781	0,821	0,719
Austria	0,884	0,847	0,744
Bahrain	0,821	0,917	0,743
Bosnia - Herzegovina	0,729	0,805	0,694
Brunei Darussalam	0,852	0,895	0,718
Canadá	0,912	0,887	0,727
Hong Kong (China)	0,908	0,833	0,824
Comoros	0,501	0,590	0,578
Dominica	0,723	0,824	0,721
Guinea Ecuatorial	0,584	0,446	0,605
Francia	0,887	0,845	0,725
Grecia	0,863	0,862	0,729
Grenada	0,742	0,801	0,713
Iraq	0,657	0,661	0,667
Irlanda	0,912	0,854	0,739
Japón	0,890	0,819	0,726
Kiribati	0,589	0,566	0,581
Kuwait	0,816	0,956	0,787
Libyan Arab Jamahiriya	0,738	0,795	0,709
Liechtenstein	0,907	0,812	0,775
Luxemburgo	0,890	0,928	0,775
Micronesia	0,639	0,596	0,623
Países Bajos	0,920	0,884	0,726
Nueva Zelanda	0,911	0,876	0,726
Oman	0,792	0,811	0,716
Palau	0,775	0,789	0,707
Portugal	0,828	0,857	0,716
Qatar	0,849	1,051	0,978
Saint Kitts and Nevis	0,747	0,828	0,703
Saint Lucia	0,729	0,837	0,704
St. Vincent & Grenadines	0,717	0,758	0,666
Samoa	0,701	0,738	0,658
Arabia Saudí	0,836	0,858	0,722
Serbia	0,771	0,821	0,717
Seychelles	0,767	0,828	0,720
Singapur	0,909	0,911	0,756
Islas Solomon	0,505	0,601	0,603
Sudán del Sur	0,461	0,435	0,554
Sudán	0,477	0,486	0,564
Timor-L este	0,601	0,466	0,604
Tonga	0,716	0,781	0,668
Turkmenistán	0,682	0,614	0,656
Emiratos Árabes Unidos	0,833	0,896	0,741

Vanuatu	0,592	0,637	0,634

La tabla de predicciones del IDH muestra valores similares obtenidos con los modelos PCR y PLS. De hecho, el estadístico RMSE vale 0,0708 para PCR y 0,0986 para PLS. Al comparar los valores predichos de IDH_2013 con los valores realmente observados, en general la predicción es razonable. Por ejemplo, en la Tabla 11, los tres países de menor valor de IDH son Afganistán, Sudán del Sur y Sudán. Los tres países con menor predicción según PLS son exactamente los mismos. El estadístico RSME sugiere que los dos modelos PCR y PLS muestran un ajuste razonable a los datos observados, pues los valores de RMSE son pequeños. No se puede asegurar si un método es mejor que otro, ya que los dos valores de RSME son similares.

La predicción del IDH para el año 2013 muestra que los modelos PCR y PLS proporcionan un ajuste razonable a los datos, los cuales permiten obtener una buena estimación del IDH en 2013 para aquellos países que no cuentan con ese dato. La predicción por medio de estos modelos basados en estructuras latentes permite aprovechar las características de estos modelos para predecir países que no cuentan con abundante información. De este modo se puede obtener una estimación para un año específico de un país, como es el caso de la predicción realizada en este trabajo, o se puede obtener una predicción promedio de un país para un conjunto de años, lo cual puede servir para conocer a grandes rasgos el nivel de desarrollo humano que tienen esos países sin disponer de la información de su IDH.

4. CONCLUSIONES

La base de datos analizada en este trabajo, obtenida por las Naciones Unidas y Banco Mundial, presentaba un alto porcentaje de datos faltantes. Por lo tanto se descartaron aquellos países y variables que no contaban con suficientes datos, obteniéndose así una matriz con 139 países y 48 variables (indicadores sociales, culturales, económicos, de salud, industria, etc.).

Los países se parecen en función de sus indicadores, los cuales están correlacionados entre sí. Por tanto, un Análisis de Componentes Principales es adecuado para entender la relación entre las variables.

China, India y Estados Unidos se identificaron como países anómalos por su distancia al modelo PCA. El motivo es que China posee valores muy altos de población total y consumo de sustancias que agotan al ozono; India por su parte tiene una población total muy elevada y altos casos de malaria, mientras que Estados Unidos presenta altas emisiones de CO_2 per cápita. Por este motivo, fueron excluidos para el resto del estudio.

Se obtuvo con PCA que la región comprendida por el sur, centro y norte de África Subsahariana está asociada a la segunda componente principal, diferenciándose del resto del mundo debido a que presenta una dinámica diferente en los valores de algunas variables, por ejemplo altos índices de enfermedades relacionadas con VIH y tuberculosis.

Exceptuando África Subsahariana, no se encontraron diferencias llamativas en cuanto a las asociaciones entre países. Sin embargo, se han identificado algunos países que se "alejan" de sus regiones geográficas como es el caso de los países Escandinavos y de los países al sur de África Subsahariana.

Se ha observado una correlación positiva entre variables de enfermedades como VIH y tuberculosis, mortalidad infantil y maternidad adolescente, las cuales muestran una relación inversa con variables consideradas de desarrollo, calidad de vida e igualdad de género.

En general, la primera componente principal explica una cantidad elevada de la varianza total de los datos, lo que indica una alta correlación entre los indicadores estudiados en este trabajo. Esta dimensión latente explica en gran medida el desarrollo de un país medido a través del índice de desarrollo humano.

Los modelos basados en estructuras latentes PCR y PLS (despliegue tipo A), presentan similares capacidades explicativas del IDH (R^2_Y = 0,95 y R^2_Y = 0,93). Los dos tipos de despliegue de la matriz tridireccional considerados para PLS han permitido obtener una interpretación complementaria de los resultados.

El modelo PLS, el cual se basa en estructuras latentes, tiene la ventaja de que permite observar la relación entre variables y países por medio de pocos gráficos, al reducir la dimensionalidad. Además permite identificar datos atípicos y aprovecha la multicolinealidad de las variables, permitiendo en un solo gráfico discutir la relación entre variables.

El modelo PLS ajustado para predecir el IDH ofrece valores distintos en función del tipo de desdoblamiento que se considere para la matriz tridireccional, permitiendo obtener un IDH

promedio (despliegue tipo A) o un IDH para cada año (despliegue tipo B). Los valores de IDH para los países cuyo valor no se ha publicado fueron estimados por medio de PCR y PLS (desdoblamiento tipo B). Estas predicciones se obtuvieron para 35 países.

La interpretación de los distintos modelos en cuanto a la predicción de la variable de desarrollo IDH permite en su conjunto una mejor comprensión de los indicadores relacionados con el grado de desarrollo de un país. Así pues, los resultados permiten discutir los criterios generales a la hora de realizar una adecuada política pública y brindar asistencia internacional focalizada en sectores que permitan mejorar el desarrollo de los países vulnerables.

En términos generales, este trabajo concuerda con la frase del Programa de las Naciones Unidas para el Desarrollo, que afirman que:

"*En las dos últimas décadas, el desarrollo humano ha avanzado considerablemente en muchos aspectos. La mayoría de las personas disfruta hoy de una vida más prolongada y más saludable y puede acceder a más años de educación, así como a una amplia gama de bienes y servicios. Incluso en países con una situación económica adversa, en general la salud y la educación han mejorado bastante. Los avances se observan no sólo en salud, educación e ingresos, sino también en la capacidad de la gente para elegir a sus líderes, influir en las decisiones públicas y compartir conocimientos".*

5. BIBLIOGRAFÍA

Banco Mundial (2004). Culture and Sustainable Development: a Framework for Action. World Bank: Washington.

Boisier, S. (1993). Desarrollo regional endógeno en Chile. ¿Utopía o necesidad? Ambiente y Desarrollo. Vol. IX-2 CIPMA: Santiago de Chile.

Boisier, S. (2001). Transformaciones Globales, Instituciones y Políticas de Desarrollo Local. Homo Sapiens: Argentina.

Borrás, F. (2014). Conferencias para la Asignatura de Evaluación de Políticas Públicas de la Maestría de Economía de la Facultad de Economía. Inédito: La Habana.

De Kostka, E. (2013). Políticas Públicas. Materiales Docentes del Diplomado en Administración Pública - Escuela Superior de Cuadros del Estado y el Gobierno de Cuba. pp. 50 -61.

Dye, T.R. (2008). Understanding Public Policies (12th ed.). Pearson Prentice Hall: New Jersey.

Elizalde, A. (2003a). Planificación Estratégica Territorial y Políticas Públicas para el Desarrollo Local. Instituto Latinoamericano y del Caribe de Planificación Económica y Social. ILPES: Santiago de Chile.

Elizalde, A. (2003b). Desarrollo humano y ética para la sustentabilidad. Revista de la Universidad Bolivariana, 2 (6).

Esbensen, K.; Geladi, P. (1987). Principal Component Analysis. Chemometric and Intelligent Laboratory Systems, 1, 41-56.

Garofoli, G. (1995). Modelli Locali di Sviluppo. Franco Angelí: Milán.

Geladi, P. and Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. Analytica Chimica Acta, 185, 1-17.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24 (6 & 7), 417–441 & 498–520.

Jolliffe, I.T. (2002). Principal Component Analysis (2nd ed). Springer-Verlag: New York.

Maggiolo, I. (2007). Políticas públicas: proceso de concertación Estado - Sociedad. Revista Venezolana de Gerencia, 12 (39).

Mesa, C. (2014). Políticas públicas y desarrollo local en Cuba: una propuesta para el debate. Estudios del Desarrollo Social en Cuba y América Latina, 2(3), 29-44.

Morales, E. (2013). El desarrollo local y su falta de financiamiento en políticas públicas en México. Revista de la Facultad de Economía, 48, 61-96.

Murga, M.A. (2006). La educación necesaria: sinergias desarrollo - educación. En: M.A. Murga (ed.), Desarrollo Local y Agenda 21: una Visión Social y Educativa (p. 190-218). Prentice-Hall: Madrid.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6, 2(11), 559–572.

Peña, D. (2002). Análisis de Datos Multivariantes. Mcgraw-Hill: Madrid.

Piñango, R. (2003). Políticas Públicas en América Latina: Teoría y Práctica. IESA: Caracas, Venezuela (1ª ed), pp. 249-277.

Repetto, F. (2001). Gestión Pública y Desarrollo Social en los Noventa: Las Trayectorias de Argentina y Chile. Prometeo: Buenos Aires, Argentina (1ª ed), 336 p.

Romer, M. (1994). The origins of endogenous growth. Journal of Economic Perspectives, 8(1), 3-22.

Rubin, D. (1987). Inference and missing data. Biometrika, 63 (3), 581-592.

Sauvy, A. (1952). Trois mondes, une planète. L'Observateur n°118 p. 14 (14 agosto 1952).

Stiglitz, J. (1998). Más instrumentos y metas más amplias desde Washington hasta Santiago. En: Seminario sobre Estabilidad y Desarrollo Económico en Costa Rica, Reformas Pendientes. Academia Centroamericana: Costa Rica.

Stiglitz, J. (2002). El Malestar en la Globalización. Taurus: Buenos Aires, Argentina.

Subirats, J. (1989). Análisis de Políticas Públicas y Eficacia de la Administración. Instituto Nacional Administración Pública: Madrid.

Vázquez-Barquero, A. (2004). Desarrollo Endógeno y Globalización. EURE, Vol. XXVI: Santiago de Chile.

Wold, H. (1966a). Nonlinear estimation by iterative least squares procedures. In: Research Papers in Statistics, David F. (ed.). John Wiley: New York, pp. 411-444.

Wold, H. (1996b). Estimation of principal components and related models by iterative least squares. In: Multivariate Analysis, Krishnaiah P.R. (ed.). Academic Press: New York, pp. 391-420.

PAGINAS WEB

Banco Mundial. (2011). Mantener los avances en medio de la inestabilidad. Recuperado de http://www.imf.org/external/datamapper/index.php

Banco Mundial. (2016). Indicadores del desarrollo mundial. Recuperado de http://databank.bancomundial.org/data/databases.aspx

Fondo Monetario Mundial. (2016). Perspectivas económicas mundiales. Recuperado de http://www.imf.org/external/datamapper/index.php

Naciones Unidas. (2010). Los objetivos del desarrollo del Milenio. Recuperado de http://www.un.org/es/aboutun/booklet/globalization.shtml

Organización de las Naciones Unidas. (2012). Departamento de estadística. Recuperado de http://data.un.org/

Programa de las Naciones Unidas para el Desarrollo. (2006). Informe sobre desarrollo humano 2006, Más allá de la escasez: poder, pobreza y la crisis mundial del agua. Recuperado de http://hdr.undp.org/sites/default/files/hdr_2006_es_completo.pdf

Programa de Naciones Unidas para el Desarrollo. (2010). Informe sobre desarrollo humano 2010, La verdadera riqueza de las naciones; caminos al desarrollo. Recuperado de http://hdr.undp.org/sites/default/files/hdr_2010_es_complete_reprint.pdf

Programa de Naciones Unidas para el Desarrollo. (2015). Índice de Desarrollo Humano. Recuperado de http://hdr.undp.org/es/content/el-%C3%ADndice-de-desarrollo-humano-idh

Programa las Naciones Unidas para el Desarrollo. (2015). Human Development Reports. Recuperado de http://hdr.undp.org/en

6. ANEXO – Variables del trabajo

Tabla 12. Lista de variables consideradas en el trabajo

VARIABLE	SECTOR
Emisiones de dióxido de carbono (CO_2), toneladas de CO_2 per cápita	Ambiental
Emisiones de dióxido de carbono (CO_2), en miles de toneladas métricas de CO_2	Ambiental
Consumo de todas las sustancias que agotan el ozono, en toneladas métricas	Ambiental
Suscripciones de telefonía fija y celular móviles por cada 100 habitantes	Calidad de vida
Usuarios de Internet por cada 100 habitantes	Calidad de vida
Proporción de la población que utiliza fuentes mejoradas de agua potable, rural	Calidad de vida
Proporción de la población que utiliza fuentes mejoradas de agua potable, total	Calidad de vida
Proporción de la población que utiliza fuentes mejoradas de agua potable, urbano	Calidad de vida
Proporción de la población que utiliza servicios de saneamiento mejorados, rural	Calidad de vida
Proporción de la población que utiliza servicios de saneamiento mejorados, total	Calidad de vida
Proporción de la población que utiliza servicios de saneamiento mejorados, urbano	Calidad de vida
Índice de Desarrollo Humano (IDH)	Calidad de vida
Servicio de deuda como porcentaje de las exportaciones de bienes y servicios y los ingresos netos	Economía
Uso de energía (kg equivalentes de petróleo) por $ 1,000 PIB (PPP $ constantes de 2005)	Economía
Formación Bruta de Capital (% del PIB)	Economía
Población total	Economía
PIB per cápita, (precios internacionales constantes 2011 $)	Economía
Crecimiento del PIB, en dólares constantes de 2010	Economía
Índice de Gini (estimación del Banco Mundial)	Economía
% de alumnos que comienzan el primer grado y llegan al último grado de primaria, ambos sexos	Educación
Tasa de finalización de la primaria, ambos sexos	Educación
Tasa total neta de matrícula en la enseñanza primaria, ambos sexos	Educación
Gasto en educación como % del gasto público total (%)	Educación
El gasto público en las instituciones educativas como % del PIB (%)	Educación
El gasto público en instituciones terciarias como % del PIB (%)	Educación
El gasto público en educación como % del PIB (%)	Educación
Esperanza de vida escolar, primario hasta el terciario, de ambos sexos (años)	Educación
Índice de paridad de género en la matrícula primaria	Género
Índice de paridad de género en la matrícula de nivel secundario	Género
Índice de paridad de género en la matrícula de nivel terciario	Género
Puestos ocupados por mujeres en el parlamento nacional, porcentaje	Género
Proporción de mujeres entre los empleados remunerados en el sector no agrícola	Género
Tasa total neta de matrícula en la enseñanza primaria, niños	Género
Tasa total neta de matrícula en la enseñanza primaria, niñas	Género
Tasa de natalidad entre las adolescentes, por cada 1.000 mujeres	Salud
Muertes por sida	Salud
Niños de 1 año vacunados contra el sarampión, porcentaje	Salud
Tasa de mortalidad de niños menores de cinco años por cada 1.000 nacidos vivos	Salud
Tasa de incidencia del VIH, 15-49 años, porcentaje	Salud
Tasa de mortalidad infantil (0-1 años) por cada 1.000 nacidos vivos	Salud
Personas que viven con el VIH, 15-49 años, porcentaje	Salud
Tasa de mortalidad de la tuberculosis por año por cada 100.000 habitantes	Salud
Tasa de detección de la tuberculosis con DOTS, porcentaje	Salud
Tasa de incidencia de tuberculosis por año por cada 100.000 habitantes	Salud
Tasa de prevalencia de la tuberculosis por cada 100.000 habitantes	Salud
Tasa de éxito del tratamiento de la tuberculosis bajo DOTS, porcentaje	Salud
Vacunación contra sarampión (% de niños entre las edades de 12-23 meses)	Salud
Casos de malaria reportados	Salud