Sáez, C.; Garcia-Gomez, JM. (2018). Kinematics of Big Biomedical Data to characterize temporal variability and seasonality of data repositories: Functional Data Analysis of data temporal evolution over non-parametric statistical manifolds. International Journal of Medical Informatics. 119:109-124. https://doi.org/10.1016/j.ijmedinf.2018.09.015

Kinematics of Big Biomedical Data to characterize Temporal Variability and Seasonality of data repositories: Functional Data Analysis of data temporal evolution over non-parametric statistical manifolds

Carlos Sáez (carsaesi@upv.es)[1] and Juan Miguel García-Gómez (juanmig@upv.es)[1]

[1] Biomedical Data Science Lab (BDSLab), Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA), Universitat Politècnica de València (UPV), Camino de Vera s/n, Valencia 46022, España

Abstract

Aim: The increasing availability of Big Biomedical Data is leading to large research data samples collected over long periods of time. We propose the analysis of the kinematics of data probability distributions over time towards the characterization of data temporal variability.

Methods: First, we propose a kinematic model based on the estimation of a continuous data temporal trajectory, using Functional Data Analysis over the embedding of a non-parametric statistical manifold which points represent data temporal batches, the Information Geometric Temporal (IGT) plot. This model allows measuring the velocity and acceleration of data changes. Next, we propose a coordinate-free method to characterize the oriented seasonality of data based on the parallelism of lagged velocity vectors of the data trajectory throughout the IGT space, the Auto-Parallelism of Velocity Vectors (APVV) and APVVmap. Finally, we automatically explain the maximum variance components of the IGT space coordinates by means of correlating data points with known temporal factors from the domain application.

Materials: Methods are evaluated on the US National Hospital Discharge Survey open dataset, consisting of 3,25M hospital discharges between 2000-2010.

Results: Seasonal and abrupt behaviours were present on the estimated multivariate and univariate data trajectories. The kinematic analysis revealed seasonal effects and punctual increments in data celerity, the latter mainly related to abrupt changes in coding. The APVV and APVVmap revealed oriented seasonal changes on data trajectories. For most variables, their distributions tended to change to the same direction at a 12-month period, with a peak of change of directionality at mid and end of the year. Diagnosis and Procedure codes also included a 9-month periodic component. Kinematics and APVV methods were able to detect seasonal effects on extreme temporal subgrouped data, such as in Procedure code, where Fourier and autocorrelation methods were not able to. The automated explanation of IGT space coordinates was consistent with the results provided by the kinematic and seasonal analysis. Coordinates received different meanings according to the trajectory trend, seasonality and abrupt changes.

Discussion: Treating data as a particle moving over time through a multidimensional probabilistic space and studying the kinematics of its trajectory has turned out to a new temporal variability methodology. Its results on the NHDS were aligned with the dataset and population descriptions found in the literature, contributing with a novel temporal variability characterization. We have demonstrated that the APVV and APVVmat are an appropriate tool for the coordinate-free and oriented analysis of trajectories or complex multivariate signals.

Conclusion: The proposed methods comprise an exploratory methodology for the characterization of data temporal variability, what may be useful for a reliable reuse of Big Biomedical Data repositories acquired over long periods of time.

Keywords: temporal stability, data quality, time series, data reuse, big data, seasonality, coordinate-free, trajectories, functional data analysis, statistical manifolds

Author version of manuscript published in the International Journal of Medical Informatics
(https://doi.org/10.1016/j.ijmedinf.2018.09.015)

1 Introduction

Big Biomedical Data repositories are increasingly available. Publicly available Open Data research repositories and property biomedical research databases, are becoming bigger both in terms of sample size and collected variables [1, 2]. Two significant reasons behind this are the widespread adoption of data-sharing initiatives and technological infrastructures, and the continuous and systematic population of those repositories over longer periods of time. However, it is acknowledged that these two situations can also introduce potential confounding factors in data which may hinder their reuse for research [3, 4, 5, 6, 7], such as in population research or in statistical and machine learning modelling. Concretely, differences in protocols, populations, or even unexpected biases, either caused by systems or humans, can lead to undesired heterogeneity in data among their sources or over time. This multi-source and temporal variability of data will be reflected on its statistical distributions, related to the above-mentioned confounding factors which, in the end, represent a Data Quality (DQ) issue which must be addressed for a reliable data reuse [6, 8].

In this work, we focus on providing a comprehensive methodology to help data-driven biomedical researchers in characterizing the temporal variability that can be present in research repositories acquired over long periods of time. In general, there is more awareness about the statistical variability that may be introduced when dealing with different data sources, such as in cross-border, multi-site repositories, when dealing with biospecimens acquired at multiple laboratories, or in clinical trials data introduced by multiple professionals. In this line, from traditional statistical univariate methods such as the ANOVA, through batch effect adjustment mechanisms [9, 10, 11], until multivariate DQ metrics [7] are generally employed to deal with multi-source variability.

Time has also received some attention as a factor of change affecting the reuse of data. However, this has been mainly studied in the domains of change detection and time series, and only a few works have related temporal variability to a DQ issue in the reuse of research biomedical data [3, 5]. Temporal variability can have a significant effect on the effectiveness and efficiency of data-driven biomedical research [6, 9]. Time in healthcare processes can also leave an imprint on electronic health records (EHR) data what is predictive to patients status of health [12]. In fact, the International Medical Informatics Association (IMIA) recently highlighted the value of temporal relationships between data, as found in their review of the literature published in 2016 regarding the Secondary Use of Patient Data [8]. Therefore, the benefits of specific temporal variability techniques can be of utmost importance in the present, but especially in the future Big Biomedical Data research.

In previous work [5], we contributed with the Information Geometric Temporal (IGT) plots to support the exploration of temporal variability of heterogeneous biomedical data, including multivariate, multimodal distributions and multiple types of variables. IGT plots project data temporal batches as a series of points where the distances among them correspond to the dissimilarity of their statistical distributions, namely a non-parametric statistical manifold [13, 14]. In this manner, the temporal relationship between the points in the projected space shows an empirical layout of data behaviour over time. The results of that work remained at the data visualization stage, but the developed technique opened the way to further possibilities for temporal variability assessment, which are now proposed in this work.

Concretely, in the present study, we aim to understand the rationale of temporal variability in terms of describing trends, abrupt changes and seasonality, the main outcomes of conventional time series analysis, but with the challenge that we are in a multidimensional non-parametric statistical manifold constructed from heterogeneous biomedical data. Concretely, considering an inherent continuous temporal flow through the projected discrete temporal batches, we estimate a continuous *data temporal trajectory* from which to study its kinematics. This estimation is made based on the well established Functional Data Analysis (FDA) technique [15]. Data kinematic properties give light to measurements about the velocity and acceleration of changes in data. The estimated trajectory allowed us constructing a novel coordinate-free (or trajectory-intrinsic) method to quantify the seasonality of data over the embedded IGT space based on the parallelism of lagged velocity vectors. Finally, to automatically provide semantics about the components of temporal variability, we propose a method to relate the IGT plot coordinates to specific temporal factors. The proposed methods have been evaluated in the large open data repository of the US National Hospital Discharge Survey [16] (3,25M hospital discharges from 2000 to 2010), contributing with a series of novel temporal variability findings.

The rest of the paper is organized as follows. Section 2 reviews related work and summarizes our

background work on temporal variability. Section 3 describe the technical development of the proposed methods. Next, the NHDS data used in the evaluation is introduced in Section 4. Section 5 describes the evaluation results related to each of the methods. In Section 6 this work is discussed in terms of its significance and implications. Some limitations are discussed too. Finally, Section 7 concludes this work and compiles its main highlights.

2 Background

This work stands as a medical informatics interdisciplinary research in the areas of Big Biomedical Data, data quality, time series, change detection and functional data analysis. Next, we describe some background work on these areas, followed by a review of the previous baseline work about IGT plot projections on temporal, non-parametric statistical manifolds.

2.1 Time in Data Quality

Data Quality is data that are fit for use [17]. DQ is characterized by DQ dimensions, as attributes that represent single aspects or constructs of DQ, which can conform to data specifications or to user expectations [17, 18, 19]. Several works have reviewed the DQ literature regarding dimensions for the reuse of biomedical data [20, 21, 22]. Among these, time is included in dimensions such as timeliness, currency or volatility. However, these dimensions are generally related to an individual data level, i.e., whether individual data registries are up-to-date compared to their real-world values, or what is their rate of change [23, 24]. We refer the reader to Table I in the work by Heinrich et al. [23] and Table II in the work by Batini et al. [24]. But, at the population level, the processes that generate data do not need to be stationary, leading to the additional issue that data subsamples are not concordant over time. This may be due to changes in protocols, in the inherent biological and social-behaviour, or even to unexpected biases caused by systematic or random errors. In clinical trials or public health registries studies, this temporal issue has been defined as the concordance or comparability dimensions over time [25, 3, 4]. In a previous work [5], we made more specific the concept of temporal concordance over time as a *temporal stability* DQ dimension.

2.2 Time series and change detection

A time series is a set of observations $\{x_t\}$, which are registered at specific times $\{t\}$ [26], with $t = 1, \ldots, T$. Generally, time series analysis is made on discrete time series, i.e., those where observations are made at discrete, equidistant time intervals. In this case, time series are more formally defined as $\{x(t)\}$, in contrast to $\{x_t\}$ which generally represents continuous observations [26]. The most common applications of time series are for univariate series, i.e., a single feature being observed over time. This single feature can correspond to an individual object being measured, e.g., a patient blood saturation level in an ICU, or it may correspond to a summary of a sample, e.g., the analysis of average incidence rates of a disease. Besides, other applications may involve the analysis of multiple time series simultaneously.

The two traditional aims of time series analysis are (1) describing information about the stochastic process generating a series and (2) forecasting. Regarding (1), as related to the purpose of this work, time series are commonly described in terms of trends and seasonality. A trend is a systematic and continuous change towards a direction (linear or not) occurring in the series, non-repetitively along the full-time period, or within a sub-period. An example of this is a yearly increase in the incidence of lung cancer in adult patients [27]. On the other hand, the seasonality of a series can be defined as the repetition of some change patterns over a fixed time period. Several methods exist to detect and characterize trends and seasonality, where some of them can be worked for both tasks.

Trends can be identified by means of fitting smoothing functions to data [28], which allow approximating a continuous parametric function to a discrete set of points while removing noise or high frequency components of change, thus revealing cleaner large-scale structures such as trends. Traditional smoothing functions include moving average, exponential smoothing, low-pass filters, b-splines or Fourier series among others.

Besides, seasonality can be formally defined as a correlational dependency of order k between the observations $x(t)$ and $x(t+k)$, where k is referred as lag. This process is generally carried out by means of the discrete autocorrelation function at a given lag k [29]:

$$r_k = \frac{\frac{1}{T}\sum_{t=1}^{T-k}(x(t)-\bar{x})(x(t+k)-\bar{x})}{\sigma_x}, \tag{1}$$

where, for a set of monotonically increasing lags $k = 1, ..., l$ versus their corresponding autocorrelations $r_1, ..., r_l$ leads to the well-known correlogram plot, which allows easily identifying seasonality at specific periods (lags).

Depending on the study purpose, seasonality is also studied by means of Fourier time-frequency transformations. A (discrete) Fourier transform of the discrete time series $\{x(t)\}$ is given by:

$$X_f = \sum_{t=0}^{T-1} x(t) \cdot e^{-i2\pi ft/T} = \sum_{t=0}^{T-1} x(t) \cdot (cos(2\pi ft/T) - i \cdot sin(2\pi ft/T)), \tag{2}$$

where X_f is a sequence of complex numbers encoding the amplitude and phase of the sinusoidally related component $e^{-i2\pi ft/T}$. Roughly speaking, the magnitude of X_f can be defined as the amount of signal with frequency $2\pi f$ in the series $\{x_t\}$. This magnitude is, then, associated to the degree of seasonality of the time series at a given period $k = \frac{1}{2\pi f}$ time units.

Fourier analysis also results useful for trend analysis when, as mentioned before, used as a low-pass filter by means of reconstructing the original series $\{x(t)\}$ with removing high-frequencies. This is achieved using the inverse Fourier transform:

$$x(t) = \frac{1}{T}\sum_{k=0}^{T-1} X_f \cdot e^{i2\pi ft/T}, \tag{3}$$

where a specific frequency can be removed by setting its corresponding X_f value to 0.

When the aim of time series analysis is to identify the specific time points at which the sufficient statistic of a sample *changes*, we enter the field known as change detection [30, 31]. In change detection, changes are characterized as gradual, abrupt and recurrent. A variety of methods exist to deal with this classification, from the classical Statistical Process Control (SPC) by Shewhart and Deming [32] and the Page Hinkley Test [33]. Change detection methods make use as well of smoothing functions to focus their detection on specific trends, from moving window to fading approaches [34, 35, 36], the latest giving more importance to recent data. In a previous work, we designed the Probability Distribution Function SPC algorithm [5], an SPC based method to monitor changes in non-parametric probability distributions of biomedical data throughout their full shape.

Conventional time series and change detection methods are designed for univariate stochastic processes. That means that, in a multivariate process, the methods above are applied individually to the different variables composing the multivariate series. In the case of smoothing, in general, this does not imply a problem since univariate smoothing leads to equally well-approximated signals. However, in the case of Fourier transforms and autocorrelations, they are designed for univariate time series, with no direct application to multidimensional measurements incorporating knowledge about variable relationships. In this case, and as done in some change detection methods, multiple variables can be summarized into an individual sufficient statistic providing the required joint information, such as using aggregates, covariates, or dimensionality reduction methods.

Finally, it is worth to mention the case where we have d (one or more) variables of the same individual being acquired over time, thus forming a d-dimensional multivariate time series $\{\mathbf{x}(t)\}$. Thereafter, $\{\mathbf{x}(t)\}$ can be defined as the trajectory of the object over the d-dimensional space on which it is defined. Here we must note that the time series methods described above will be completely dependent on the arrangement of the d dimensions. In other words, while applying a rotation on the trajectory would yield an exact trajectory arrangement, the results of time series analysis on its individual dimensions (such as autocorrelation, Fourier or trend analysis) would yield different results. This may happen in cases where dimensions are meaningless, or with no meaning being originally assigned, such as in the case of Information Geometric Temporal plots, as we will see next. In these cases, there is a need to establish coordinate-free time series methods.

2.3 Functional Data Analysis

The rationale behind Functional Data Analysis (FDA) is to consider a time series as a single functional entity, rather than a sequence of individual observations [15]. This leads to a wide range of techniques of great usefulness in biomedical applications including smoothing, alignment, functional principal component analysis and regression [37], but also to the analysis of continuous derivatives of an originally discrete time series, allowing the study of series kinematics.

Let define X as a functional observation, where X is the intrinsic structure of a time series consisting on T observations $\{x(t)\}$[1]. The link between X and $\{x(t)\}$ is given by

$$x(t) = X(t) + \epsilon(t), \tag{4}$$

where, $\epsilon(t)$ corresponds to the observational error or noise of observation $x(t)$ over the *true* function value $X(t)$. Herein, FDA allows modelling a time series by filtering out $\epsilon(t)$ to a desired degree, namely smoothing.

The general solution of FDA to estimate a functional X from a set of observations $\{x(t)\}$ is a linear combination of K basis functions $\phi_1 \dots \phi_K$, conforming a basis expansion, as follows:

$$X(t) = \sum_{k=1}^{K} c_k \phi_k(t), \tag{5}$$

where the number K of basis functions is related to the degree of smoothing, or approximation of X to x: a large K will overfit data in x, in contrast, a small K will yield a strong smoothing.

The problem of estimation can be simply reduced to estimate the coefficients $c_1 \dots c_K$ from the raw time series data by minimizing the least squares criterion

$$SSE = \sum_{t=1}^{T} \left[x(t) - \sum_{k=1}^{K} c_k \phi_k(t) \right]^2. \tag{6}$$

where the solution for the coefficients vector $\hat{\mathbf{c}}$ in matrix form is given by:

$$\hat{\mathbf{c}} = (\Phi'\Phi)^{-1}\Phi'\mathbf{x}. \tag{7}$$

The basis functions conforming an expansion can be chosen from families of derivable basis functions, with different properties suitable to different tasks. The most common basis functions are the Fourier basis and polynomial splines basis [15, 37], for their great flexibility to approximate most series at different levels of smoothing, and which result useful to periodic series the former, and to have local support the latter.

As an example, the Fourier basis expansion is defined as follows:

$$X(t) = c_0\phi_0(t) + c_1\phi_1(t) + c_2\phi_2(t) + \dots \begin{cases} \phi_0(t) = 1, \\ \phi_{2k-1}(t) = \sin(k\frac{2\pi}{\mathcal{T}}t), \\ \phi_{2k}(t) = \cos(k\frac{2\pi}{\mathcal{T}}t), \end{cases} \tag{8}$$

where k indicates the basis number, and $\mathcal{T}$ is related to the signal period. The period $\mathcal{T}$ is generally set as the length of the interval $T : [t_0, \dots, t_T]$ of the observed time series being smoothed. Therefore, the estimated function $X(0)$ will be periodic at that period, meaning that $X(0) \approx X(T)$. We may have reasons to use Fourier basis expansions in non-periodic series, such as when we are aware of inherent seasonal components at the series, or when we require a good approximation at the signal extremes, what will be more difficult using splines basis. In that case, to avoid the periodic behaviour of the resultant function at the original observed interval $[t_0, \dots, t_T]$, we can extend the period $\mathcal{T}$ at both extremes as $[t_0 - a, \dots, t_T + a]$, where a would normally be a fraction of T, and then get the resultant smoothed signal just by evaluating $X(t)$ between the original time points $[t_0, \dots, t_T]$.

[1] Other FDA related works describe the model for a set of functional observations $\{X_i\}$, e.g., when studying a set of time series each for a different subject. In this work, we focus on the case of a single functional observation X.

2.4 Information Geometric Temporal plots of temporal statistical manifolds

Information Geometric Temporal (IGT) plots project a number of points each related to a data temporal batch, where the Euclidean distances among them correspond to the dissimilarity of their statistical distributions [5, 6]. In this manner, the temporal relationship between the points in the projected space shows an empirical layout of data behaviour over time. To facilitate the visual exploration, time batches are labelled according to their date and coloured according to their season. IGT plots are based on the concept of non-parametric Information Geometry, by which probability distributions lie on a Riemannian manifold which metric space can be defined by Kullback-Leibler derived divergences, namely a non-parametric statistical manifold [13, 14]. Overall, the method is not restricted to the use of other distribution distances if needed.

Let $\mathbf{R}$ be a data repository, which is subdivided in T batches $\mathbf{R}(1), \ldots, \mathbf{R}(T)$ in a periodic timely basis, e.g., monthly. For each batch, we estimate the probability distribution of its data P_t. Depending on the purpose, we can estimate the distribution of specific variables (univariate), the joint distribution of several variables (multivariate), or the distribution of a dimensionally reduced version of the dataset. Several approaches can be used to estimate the distributions according to the types of involved variables, such as histograms for categorical or short-range numerical variables, or Kernel Density Estimations [38] as a general method with continuous output for numerical data. Now, for the sequence of distributions $P(1), \ldots, P(T)$ we calculate the $\binom{T}{2}$ pairwise distances, such as using the Jensen-Shannon distance [39, 40]:

$$JSD(P_i||P_j) = JS(P_i||P_j)^{1/2} = \left(\frac{1}{2}KL(P_i||M) + \frac{1}{2}KL(P_j||M)\right)^{1/2} \tag{9}$$

where $M = \frac{1}{2}(P_i + P_j)$, and $KL(P||Q)$ is the Kullback-Leibler divergence between distributions P and Q [41]. These distances can be arranged in a T-by-T symmetric dissimilarity matrix $Y = (y_{11}, ..., y_{TT}), y_{ij} : JSD(P_i||P_j)$, which is then used as the input to multidimensional scaling (MDS) [42]. The objective of MDS is to obtain the set $\mathbf{x}_1, ..., \mathbf{x}_T$ of points in a d-dimensional Euclidean space by finding the best approximation of $||\mathbf{x}_i - \mathbf{x}_j|| \approx y_{ij}$, using specific loss functions [43]. Supported by eigendecompositions, likewise Principal Component Analysis, MDS sorts the output dimensions according to the amount of variance in the cloud of projected points. The output of MDS is then used to project and visualize (for $d \in [1, 3]$) the IGT plot, where a point p_t is related to the distribution of the dataset time batch R_t.

Next, we define some conceptual terms with respect to the analysis of temporal variability on IGT plots, which investigation towards their characterization forms the aim of this work.

Definition 1. ***IGT space.*** *Space at which the projected temporal batches of the IGT plot lie. Any semantic interpretation of the projection comes from the relativeness among the positions of time batches. According to the MDS projection, the axis coordinates of the IGT space do not have a specific meaning but the components of variance. The origin of coordinates* $\mathbf{O}$ *does not have any semantic meaning except that time batches are centred on it. One of the aims of this work is to investigate if, assuming that variance is not related to randomness, we can assign a meaning to the IGT space coordinates.*

Definition 2. ***Trend.*** *Continuous and smooth change in the probability distributions of time batches over time, along the full-time period, or within a sub-period. Trends can be linear or, more generally, curved. Trends are represented in the IGT space as a continuous flow of time batches through a direction related to time.*

Definition 3. ***Abrupt change.*** *A sudden change in probability distributions at a specific time point, leading to a new data inherent concept which is maintained afterwards. Abrupt changes are represented in the IGT space as a gap between two groups of continuous time batches. Multiple abrupt changes can occur in a data repository, splitting the dataset into multiple clusters of time batches (see the definition of temporal subgroups). A single time batch could be abruptly separated from the rest, generally due to some specific context in its data (e.g., transient states, incomplete batches), in that case, we will talk about an outlier batch.*

Definition 4. ***Temporal subgroups.*** *Conceptually related groups of time periods at which probability distributions are similar within a group, but dissimilar between groups, i.e., forming clusters of time batches. Abrupt changes do generally split data into temporal subgroups. A consecutive time flow between*

batches at two temporal subgroups would indicate a recurrent behaviour. An outlier batch will not be considered within any subgroup.

Definition 5. ***Seasonality*** *Repetition of some change patterns at a specific time period throughout the IGT space. Seasonality should be represented in the IGT space as repetitive cycles over the general temporal flow. We could find local seasonality, within a specific time period, or global seasonality, along the full study period. Global seasonality should be maintained even across multiple temporal subgroups, e.g., in a data repository which is partitioned on various temporal subgroups, a global yearly variation should be maintained across the different subgroups.*

3 Proposed methods

The proposed methods are divided into two groups. The first one consists of two methods to model and describe the kinematic behaviour of Big Data repositories throughout their IGT spaces. These methods allow measuring and characterizing trends, abrupt changes and seasonality on IGT spaces. These methods are supported by the assumption that, despite the discrete nature of time batching, there exist an inherent continuous structure on data evolution, i.e., data can be considered as a particle moving over time through a multidimensional probabilistic space. This is driven into practice by modelling a continuous trajectory of the data motion through their IGT spaces using FDA.

Next, we propose a third method to complement with semantics the data changes described by the former model. This is done by assigning a meaning to the axis coordinates of the IGT spaces, by means of correlating data points with known temporal factors from the domain application, leading to an interpretation of the data motion vectors.

3.1 Kinematic model

Let the sequence of points $\{\mathbf{x}(t)\} = \mathbf{x}(1), ..., \mathbf{x}(T)$ in an d-dimensional IGT space, where $\mathbf{x}(t)$ represents the distribution $P(t)$ of a data batch $\mathbf{R}(t)$ at time t. The application of FDA (Section 2.3) allows us to estimate a continuous function for $\{\mathbf{x}(t)\}$ in terms of time: $\mathbf{X}(t)$ (Equation 5). This function provides us with a smoothed, derivable trajectory of $\{\mathbf{x}(t)\}$, describing the underlying kinematic behaviour of data over time, and leading to the calculus of the following motion equations on IGT spaces:

$$\mathbf{v}(t) = \frac{d\mathbf{X}(t)}{dt}, \tag{10}$$

where $\mathbf{v}(t)$ is the velocity vector of the distribution $P(t)$ through the IGT space at time t,

$$\mathbf{a}(t) = \frac{d^2\mathbf{X}(t)}{dt^2} \tag{11}$$

where $\mathbf{a}(t)$ is the acceleration vector of the distribution $P(t)$ through the IGT space at time t.

The velocity vectors indicate the direction of change of data at specific time points in the IGT space. Further, celerity, the modulus of the velocity vector $\|\mathbf{v}(t)\|$, provides us with the magnitude of change in data at a specific time point. On the other side, the acceleration vectors indicate the direction of the tendency of change of data at specific time points. Equivalently, the modulus of the acceleration vector, $\|\mathbf{a}(t)\|$, provides us with the magnitude of that tendency of change, namely the potential of change of data at specific time points.

Given the continuous and derivable structure of X, we can get instant calculations of $\mathbf{v}$ and $\mathbf{a}$, throughout the full-time period, rather than at the original specific discrete times. Therefore, if data batch acquisition is on a monthly basis, we will be able to calculate the velocity and acceleration vectors even on a daily basis. This allows us to monitor velocity and acceleration over time, where a fixed velocity over time may be an indicator of a trend, while a rapid increase in velocity followed by a decrease, what will be well reflected in acceleration, may be an indicator of an abrupt change. Additionally, periodicity in velocity and acceleration may also be a sign of seasonality, although with changes to be further explained (e.g., if, although periodic, acceleration is on random directions).

Either way, a significant outcome of this kinematic model is the capability of interpreting the directions of these motion vectors. As described in Definition 1, the vector directions have a direct interpretation

in terms of the relativeness of the points in the IGT space. The distribution at time t changes to that distribution that would be expected at the point where its velocity and acceleration vectors point out. This targeted distribution can be obtained as a weighted average of its surrounding real distributions, given the non-parametric nature of the statistical manifold (we could obtain an exact point-to-distribution map in a parametric Information Geometric domain, but that is out of the scope of this work). However, this point-relative interpretation can be complex, as traditionally we would expect a meaning of the axis coordinates. In this regard, we will describe next in Section 3.3 a method to complement this kinematic model with the interpretation of the axis coordinates in the IGT space.

Besides the meaning of coordinates, we can define a unit of measurement for any point-wise distance or vector magnitude on the IGT space. As a premise, the unit of measurement of the pairwise distance between probability distributions used in the process of creating the IGT space (Section 2.4) will define the unit of measurement in such a space. Formally, such metrics in the IGT space are an approximation, since the manifold embedding into lower dimensions can entail a loss of precision with respect to the input dissimilarity matrix, however, in our tests this loss was minimum so we consider the same metric in the embedded space. Therefore, being the pairwise distance the Jensen-Shannon distance (Equation 9), the magnitude of a velocity vector $\|\mathbf{v}\|$ could be given, e.g., in Jensen-Shannon units per month ($[JSD]/m$), where the JSD can be interpreted as a percentage of dissimilarity between two distributions.

Another property of this model is that we can apply time-frequency analysis, such as described in Section 2.2, at a higher resolution than using the original discrete points, and in a smoother signal. We can apply the Fourier transform (Equation 2) to the individual components of the trajectory X, and to the scalars celerity $\|\mathbf{v}(t)\|$ and potential of change $\|\mathbf{a}(t)\|$ to obtain their frequency representations. Then, by means of the inverse Fourier transform (Equation 3) we could be able to reconstruct the trajectory X while removing any seasonal behaviour in order to facilitate the analysis of non-seasonal trends.

As a final remark on the kinematic model, it is not the purpose of this work to study the best basis functions for the FDA smoothing, but aiming to obtain good derivatives throughout the full signal, we have chosen the Fourier basis (Equation 8) for two reasons: (1) based on periodic components they can provide a good modelling of seasonal behaviour and the basis coefficients can be interpreted as the weight of specific periods, (2) the smoothed signal provide excellent behaviour at their extremes, in contrast to b-splines which can result unstable (although in this case the coefficients will lose their interpretation).

3.2 Auto-Parallelism of Velocity Vectors

Let X be a continuous time series trajectory in a multidimensional space, i.e., the data trajectory X in the IGT space provided by the kinematic model. We can obtain the magnitude of trajectory orientation dependency of order k in the series as the degree of parallelism of k-spaced velocity vectors $\mathbf{v}(t)$ and $\mathbf{v}(t+k)$. With a functional structure similar to autocorrelation (Equation 1), our multidimensional and trajectory-oriented seasonality function is given by the following Auto-Parallelism of Velocity Vectors (APVV) equation:

$$APVV_k = \frac{1}{T-k} \sum_{t=1}^{T-k} 1 - \frac{\angle\, \mathbf{v}(t)\ \mathbf{v}(t+k)}{\pi}, \tag{12}$$

where k correspond to the query lag, and $1 - \frac{\angle\, \mathbf{v_i}\ \mathbf{v_j}}{\pi}$ corresponds to the angular similarity between velocity vectors, where $\angle$ denotes the angle between two vectors, and, as normalized by π, the similarity is bounded between 0, for exactly opposite vectors, and 1, for exact vectors.

The $APVV_k$ measurement can be supported with confidence intervals. For a given lag k, the 95% CI for an $APVV_k$ can be obtained as the $(.025, .975)$ percentiles of the statistical distributions generated from the $T-k$ k-spaced angular similarity measurements. As a result, we can generate a correlogram-like plot of the set of $APVV_k$ measurements and their confidence intervals $(APVV_{k,l}, APVV_{k,u})$ for a set of lags $k \in [1, l]$.

The APVV method above gives us a general measurement of the multivariate, oriented, and seasonality of order k over the full series through an aggregated measure consecutive of k-spaced velocity vectors. However, within a given k, starting from two random time points t_i and t_j, their corresponding pairs of k-spaced vectors $[\mathbf{v}(t_i), \mathbf{v}(t_i+k)]$ and $[\mathbf{v}(t_j), \mathbf{v}(t_j+k)]$ can be parallel intra-pair, but not inter-pair.

In other words, further than a positive k-dependency on a time series, we could also measure whether the orientation remains parallel throughout the full series at equally k-spaced vectors, and at different starting points. This can provide additional semantics to seasonal orientability. E.g., in a monthly basis, for $k = 12$ we could measure whether there is a general data velocity orientation at a specific month of the year. To this end, we complement the APVV method with the $APPV_{map}$ measuring the degree of parallelism of the series for the different lags k and starting points t, as follows:

$$APVVmap_{k,t} = f_{parall}(\mathbf{v}(t), \mathbf{v}(t+1k), \mathbf{v}(t+2k), \ldots, \mathbf{v}(t+nk)), \tag{13}$$

where for each lag k and a given starting point t, we measure the degree of parallelism among all the k-spaced velocity vectors throughout the series, and $\mathbf{v}(t+nk)$ is the last incremented velocity vector within the series length. We use the f_{parall} function:

$$f_{parall}(\mathbf{v_1}, \ldots, \mathbf{v_n}) = \frac{1}{C_{n,2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n-1} 1 - \frac{\angle\ \mathbf{v_i}\ \mathbf{v_j}}{\pi}, \tag{14}$$

where $C_{n,2}$ is the binomial coefficient $\binom{n}{2}$ averaging the total sum of pairwise similarities. Therefore, if the velocity vectors within a k are mostly parallel and point to a common direction, this indicates that there is a common direction of change, so that there exists an oriented seasonality at a period k. This will be reflected in an $APPV_{map}$ matrix, with the starting points t as rows and lags k as columns. Note this will be an upper triangular matrix, since starting from $t = k + 1$ the starting point would have been already included in $t = 1$, thus repeating the results. Also, note that for large k the number of included k-spaced velocity vectors may be low, thus causing noisy results. For this reason, the results of $APPV_{map}$ should be limited to lower values of k.

We would like to highlight that, thanks to the use of angle-based functions, these methods are coordinate-free, that is, their results will be exactly the same against rotations of the trajectory time series in its space. This results in an obvious advantage against applying traditional seasonality methods, such as the autocorrelation or Fourier analysis, on the different dimensions of the series. Additionally, it is able to detect seasonal multivariate interactions based on the directionality of the time series velocity vectors.

3.3 Automatic explanation of principal temporal variability components

Let an IGT space defined by the set of dimensions $\mathcal{D}_1, \ldots, \mathcal{D}_D$, where D is the number of dimensions chosen for the non-parametric statistical manifold embedding (Section 2.4). Since the embedding is based on eigendecomposition through classical MDS, the coordinate axes are ordered in terms of explained variance, but they have no direct meaning. The following method quantifies the contribution of different temporal factors envisaged to explain the temporal variance on each coordinate axis. This is done by correlating, for each individual coordinate in the IGT space, the points $\{\mathbf{x}(t)\} = \{[x_1(t), x_2(t), ..., x_D(t)]\}$ of the distribution trajectory throughout the space with a set of candidate explanatory factors, described next, obtaining for each pair dimension-factor a contribution level $r_{F_f,\mathcal{D}_d}$. To capture the non-linear nature of the IGT projection embeddings, we will use and denote as $corr(\{x\}, \{y\})$ the Spearman's correlation between series $\{x\}$ and $\{y\}$. Besides, we will denote as $autocorr(\{x\}, k)$ the autocorrelation of series $\{x\}$ at lag k. In this work we study the following three explanatory factors:

F1-Date The passage of time, a factor generally related to smooth changes, such as those expected by social or population trends (e.g., ageing). This is calculated by correlating the trajectory points at given dimension $\{x_d(t)\}$ with their respective dates in a serial format (e.g., in the Matlab software, as the number of days that has passed since January 0, 0000), as follows:

$$r_{F1,\mathcal{D}_d} = corr(\{x_d(t)\}, \{t\}). \tag{15}$$

F2-Seasonality A specific period at which the distribution shows a seasonal behaviour, e.g. 12 months as a yearly periodic change. This is calculated as the autocorrelation coefficient (Equation 1) of the trajectory points at given dimension $\{x_d(t)\}$ at the lag given by the desired temporal period (e.g., in a monthly basis data, for a 12-month period query then *k=12*), as follows:

$$r_{F2,\mathcal{D}_d} = autocorr(\{x_d(t)\}, k). \tag{16}$$

F3-Temporal subgroups The temporal subgroups sorted by time. The aim is to measure the contribution to dataset variability caused by these subgroups and their causing abrupt changes. Temporal subgroups can be automatically determined by means of a clustering algorithm on the discrete points of the IGT plot. This will allow labelling time batches with the corresponding temporal subgroup at which they belong. Therefore, for a set of S temporal subgroups ordered by time such as $\{s_1, s_2, ..., s_S\}$, where $\forall \mathbf{x}(t) : \mathbf{x}(t) \in s_s$, we obtain the subgroup centroids $\{\mathbf{x}_{s_1}, \mathbf{x}_{s_s}, ..., \mathbf{x}_{s_S}\}$. Next, we get the new d-dimensional series $\boldsymbol{\mathcal{C}}(t)$ relating each point $\mathbf{x}(t)$ with the centroid of its corresponding subgroup at a specific dimension d. The contribution level of F3 is then calculated by correlating $\{x_d(t)\}$ with the subgroup centroids at that dimension $\{\mathcal{C}_d(t)\}$, as follows:

$$r_{F3,\mathcal{D}_d} = corr(\{x_d(t)\}, \{\mathcal{C}_d(t)\}). \tag{17}$$

As a result of the above, we can obtain the $\boldsymbol{\mathcal{R}}$ matrix relating the dimension-factor contribution levels quantifying the variability semantics behind each dimension of the IGT space. Based on correlations, measurements are given between -1, negative correlation, and 1, positive correlation, where 0 means the dimension is independent of the factor. Normally, two or three dimensions will be used with the purpose of visualization. For instance, for the three factors above and $D = 3$, the matrix $\boldsymbol{\mathcal{R}}$ will provide the following information:

$$\boldsymbol{\mathcal{R}} = \begin{pmatrix} r_{F1,\mathcal{D}_1} & r_{F1,\mathcal{D}_2} & r_{F1,\mathcal{D}_3} \\ r_{F2,\mathcal{D}_1} & r_{F2,\mathcal{D}_2} & r_{F2,\mathcal{D}_3} \\ r_{F3,\mathcal{D}_1} & r_{F3,\mathcal{D}_2} & r_{F3,\mathcal{D}_3} \end{pmatrix} \tag{18}$$

4 Data

The methods have been evaluated in the publicly available US National Hospital Discharge Survey (NHDS) dataset [16], including a total of 3,257,718 hospital discharges between 2000 and 2010. Collected annually by the US National Center for Health Statistics, the NHDS cohort collects medical and demographic information from a sample of inpatient discharge records selected from a US national probability sample of non-Federal, short-stay hospitals (e.g., in the 2005 batch sample consisted of 444 hospitals).

The evaluation was performed based on a monthly basis granularity, leading to a total of 131 time batches. For the purpose of this evaluation, we included a heterogeneous set of numerical, categorical and coded variables, as described in Table 1. Diagnoses and procedures are coded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM)[44]. Both diagnosis and procedure codes are usually recorded in the order they were listed in the patient abstract. Besides, the maximum number of registered codes varies according to the NHDS year. However, for the purpose of evaluating the proposed methods, we focused only on the primary ones. Finally, the monthly statistical distributions of the numerical variables Age and Days of care were calculated using a Kernel Density Estimation smoothing with Gaussian kernels.

Variable	Description	Type
Age	Age in years at discharge. Ages under 1 year originally recorded at a monthly or daily level were re-coded as 0.	Numerical integer
Sex	Sex of the person.	Categorical
Days of care	Number of days of care prior to discharge.	Numerical integer
Diagnosis code #1	Primary diagnosis code, as listed in the patient abstract.	ICD-9 code
Procedure code #1	Primary procedure code, as listed in the patient abstract.	ICD-9 code
DRG	The Diagnosis-Related Group, obtained using Grouper Programs of the Centers for Medicare and Medicaid Services.	Group code

Table 1: Variables of NHDS dataset included in this study.

5 Results

In this section, we describe the results of applying the proposed methods to the NHDS data described above. Methods were applied first in a multivariate manner using all the variables in Table 1. We then

applied a univariate analysis towards a detailed explanation of the underlying temporal changes. For each case, we first describe the general kinematic behaviour of data based on the IGT plots and the kinematic figures of the data trajectory over the IGT space. Next, we describe the seasonality analysis based on the application of our multidimensional APVV and $APVV_{mat}$ methods, which are supported and compared to traditional Fourier and autocorrelation analysis on the individual dimensions of the IGT space and kinematic figures. Lastly, we describe the results of the automatic calculus of the potential explanations of temporal components of change and relate them to the previous outcomes. Additional figures in the Supplemental material file are referenced as SM-(Figure number).

We highlight that despite IGT plots and smoothed signal trajectories are visualized in 3D, their corresponding kinematic figures and APVV methods are calculated on the full, 131 dimension trajectory (based on the full-dimensional embedding of the 132 time batches). This shows the scalability of the methods to high dimensional trajectories. Comparisons between 3D and 131D results can be found in the Supplemental material.

5.1 Multivariate case

5.1.1 Description of kinematics

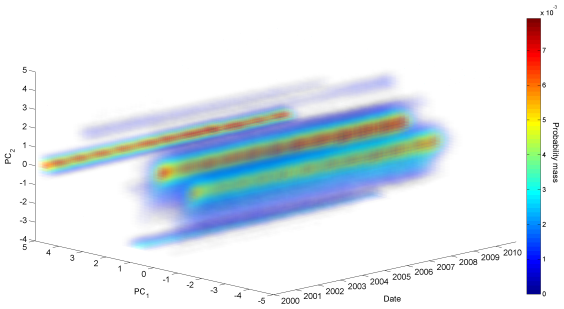
As a starting visual reference, Figure 1(a) shows a 2D probability distribution temporal heatmap of the two first PCA components of the NHDS data (each slice on the Date axis would be a 2D heatmap of the probability mass at the corresponding month). We can observe, first, that several patient clusters exist, the central ones with a higher probability mass than the surrounding ones, and in general, the clusters are maintained over time. Next, we observe that by the last fourth of the time period the two central clusters get a higher degree of probability. Also, the low-probability upper cluster seems to decrease its probability in favour of a new cluster (just next to it) which is formed by the last fourth of the time period. Finally, in the three central clusters with higher probability, we observe a cyclic behaviour over time in their central higher amount of probability mass. Although this was a rough description of data changes, we will measure the kinematics behind all these changes in the following results.

Figure 1(b) shows the IGT plot of the NHDS data, including the estimated continuous trajectory of data $\mathbf{X}(t)$. We mainly observe a cyclic seasonal behaviour, a temporal displacement over months on the D1 dimension, and an abrupt change in 2008 splitting data into two temporal subgroups. Next, Figure 1(d) shows the kinematic figures of $\mathbf{X}(t)$, i.e.: its celerity and acceleration modulus over time. We observe periodic velocity changes, with peaks twice a year, what may be related to the cyclic behaviour in the IGT plot. We also observe an overall increase in velocity in 2008, what seems related to the abrupt change in 2008.

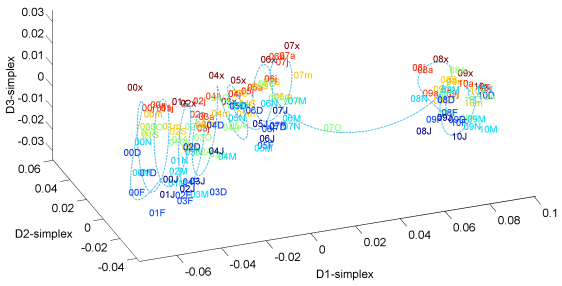
5.1.2 Seasonality

As a baseline for the results of the APVV analysis, the Fourier and autocorrelation analysis of the individual IGT dimensions in Figure 1 revealed that the cyclic behaviour in $\mathbf{X}(t)$ was present in the 2nd and 3rd dimension, likely showing a rotation behaviour over the tendency given by the 1st dimension. The Fourier analysis of the kinematic figures in Figure 1(d) confirmed the period of velocity and acceleration at 6 months.

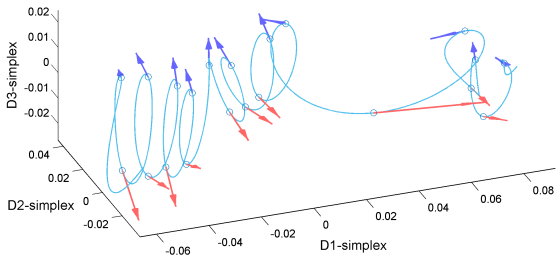
The APVV analysis in Figure 1(e) reveals a parallelism of the trajectory velocity vectors at a lag of 12 months and its respective multiples. This is confirmed by the Fourier and autocorrelation analysis of the individual IGT dimensions (Figure SM-1), revealing such a cyclic behaviour in the 2nd and 3rd coordinates of the IGT space. Additionally, the autocorrelation plot reveals that at these coordinates the signal is negatively autocorrelated at a lag of 6 months (and respective multiples). In our case, the APVV also highlights such negative autocorrelation, but in this case, the interpretation is multivariate and, particularly, it is associated to an anti-Parallelism found at the trajectory velocity vectors at a lag of 6 months (and respective multiples). These results go further than the frequencies revealed by conventional Fourier analysis or the autocorrelation on IGT coordinates and kinematic figures since they indicate that at a period of 12 months the change in data distributions is oriented towards the same direction, while at a period of 6 months the distribution change direction is opposite. This is visually demonstrated in Figure 1(c), where purple and red arrows are equispaced at 6 months, thus being parallel within colour (12-month difference) but anti-parallel between colours.



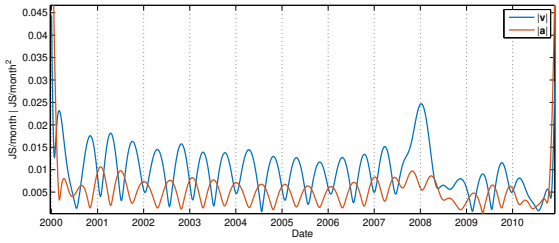
(a) 2D Probability Mass Temporal Heatmap of 2 first PCA components of NHDS data



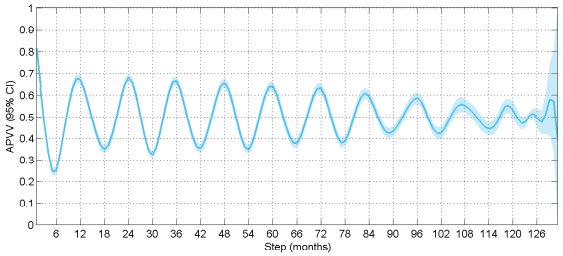
(b) Information Geometric Temporal (IGT) plot



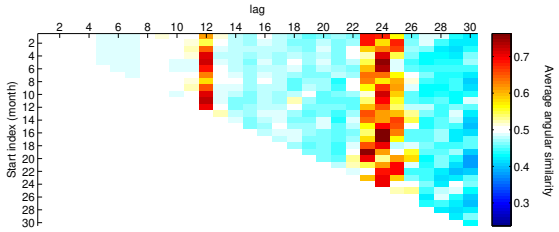
(c) 6-month spaced velocity vectors



(d) Kinematic figures of celerity $\|\mathbf{v}\|$ and acceleration $\|\mathbf{a}\|$



(e) Auto-Parallelism of Velocity Vectors



(f) APVVmap

Figure 1: Results for the multivariate NHDS data.

5.2 Univariate case

The multivariate results showed a general, wide view of the dataset temporal variability. Next, following our proposed temporal variability methodology [6] we focus on the univariate details which can give further light on the variability causes. For simplification, while ensuring that all types of findings are covered, the univariate results for kinematics and seasonality are given only for the variables Age, Days of care, Diagnosis code #1 and Procedure code #1.

5.2.1 Description of kinematics

Figure 2 shows the IGT plots with estimated signal trajectories (left) and their corresponding kinematic figures (right). Along the variables, we observe different patterns in the behaviour of data trajectories over the IGT space. For Age and Diagnosis Code #1 there is a clear seasonal behaviour, observed visually in the IGT plots according to monthly colouring and loops and supported by the periodic changes in velocity. Although this apparent seasonality cannot be clearly observed by IGT plots for Days of care and Procedure code #1, their kinematic figures show some insights, which should be checked with the specific seasonality methods next.

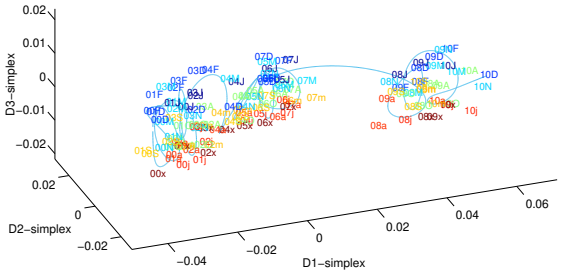
A partition into two temporal subgroups is clearly observed in the IGT plot of Age. This is supported by the sudden increase in velocity at the start of 2008 shown in the kinematic figures. This change

is related to the change of codification in Age, as described in previous works [5]. However, it results even more interesting the temporal subgroups in Procedure code #1 at a yearly basis, indicating similar procedures applied within a specific year, but different with respect to the year before and also to the coming ones. Still in Procedure code #1, the subgroup change in 2006 accounts with a magnitude of about $0.04_{\mathrm{JS/month}}$, showing a tendency of change of $\approx 4\%$ in the statistical distribution of procedures (i.e. in the histogram). Similar magnitudes of change are observed in the velocity of Diagnosis codes as well but on a yearly basis. However, in this case, this may be expected due to the evident seasonality of diseases. For additional information about the causes of the detected temporal subgroups, we refer to the Figures SM-4, SM-5 and SM-6. For Diagnosis code (Figure SM-5(b)) we observe, e.g., several ICD-9 codes which stop being used (276.5, *Volume depletion disorder*, in 2006, V27.9, *Outcome of delivery, unspecified outcome of delivery*, in 2010), which start being used (493.92, *Asthma, unspecified type, with (acute) exacerbation*, in 2001), which decrease their frequency of use (V30.00, *Single liveborn, born in hospital, delivered without mention of cesarean section*, V27.0, *Outcome of delivery, single liveborn*, and 428.0, *Congestive heart failure, unspecified*, in 2008, V57.89, *Care involving other specified rehabilitation procedure*, in 2006), which increase their frequency of use (491.21, *Obstructive chronic bronchitis with (acute) exacerbation*, in 2008, 038.9, *Unspecified septicemia*, in 2006, 518.81, *Acute respiratory failure*, in 2005, 584.9, *Acute kidney failure, unspecified*, in 2007), and several of them with a clear seasonal behaviour (e.g, 486, *Pneumonia, organism unspecified*, 414.01, *Coronary atherosclerosis of native coronary artery*, V27.0, V30.00, etc). For Procedure code (Figure SM-5(b)) we mainly observe a high prevalence of unspecified codes, as NAs value (*Not Available*), with some apparent seasonality. But here we observe several changes of ICD-9 coding which are apparently more abrupt than the in Diagnosis code, e.g., several codes stop being used (36.01, *Single vessel percutaneous transluminal coronary angioplasty [PTCA]*, in 2005, 684, *Total abdominal hysterectomy*, in 2006, 736, *Episiotomy*, in 2007), others which start being used (00.66, *Percutaneous transluminal coronary angioplasty [PTCA]*, in 2006), or others with several abrupt changes in frequencies (995.5, *Prophylactic administration of vaccine against other diseases*, 815.4, *Total knee replacement*, 736). Some of these changes in the use of codes can be related to the updates in ICD-9 coding (such as a switch from 36.01 to 00.66), others may be to the centres and regions contributing with data to the NHDS (Figure SM-12). A detailed study of these would require further investigation on specific clinical and management protocol changes, or coding changes, such as by deeply inspecting the yearly NHDS reports [16].

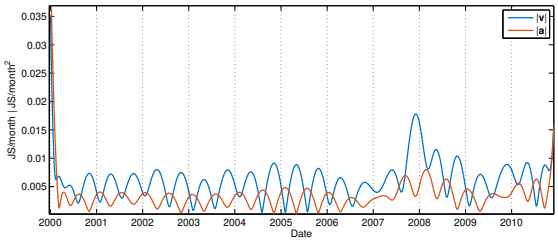
In the Days of care we initially observe a random behaviour in the IGT plot, however, the kinematic figures give signs about a possible seasonal behaviour. Nevertheless, the Days of care results required further analysis, related to the methodological settings of this analysis. By default, we set the range for numerical variables in 100 bins between the minimum and the maximum value. Considering that ages below one year were re-coded to 0 in our study to be consistent with the integer convention on older ages, this was not a problem for age (although more detail could be obtained for patients younger than one year). However, since the Days of care generally follow an exponential distribution with nearly all the density concentrated below two weeks of stay, several high length hospital stays establishing higher maximum values (larger than one year) may cause losing resolution at lower values where information detail is expected, i.e. below two weeks of stay. Therefore, we ran two additional experiments for Days of care variable by fixing the distribution bins: on 31 bins between [1,31], and on 561 bins between [1,561] (561 was the maximum observed value). The results of these are shown in Figures SM-9, SM-10, SM-11. We observed more detail such as explicit partitions into temporal subgroups, a seasonal behaviour, and in the [1,561] case we observed a clear outlier month at November 2009, caused by several large lengths of stay at that month.

5.2.2 Seasonality

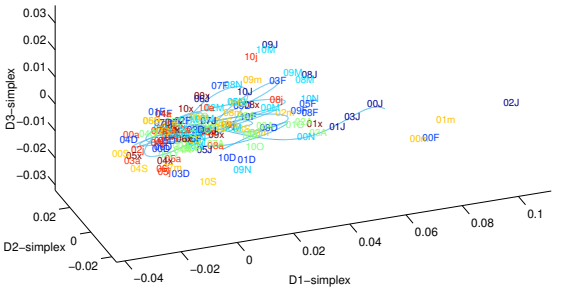
The Fourier and autocorrelation analysis at individual IGT coordinates of Age, Days of Care and Diagnosis code (Figures 3 and 4, left columns) revealed a 12-month period seasonal behaviour. This was evidenced visually in IGT plots for Age and Diagnosis code, but it was not as clear for Days of care. Interestingly, we observe that the main coordinates in that periodic behaviour differ among the three variables. For instance, in days of care, the peak is in the first dimension, that containing the largest variance on the IGT projection. Thus, we could consider that the seasonality is highly contributing to the temporal variance of data for that variable. These 12-month periodic patterns agree with the 6-month period on velocities (Figure 3, right column). We also find a high peak in speed and acceleration at a period of



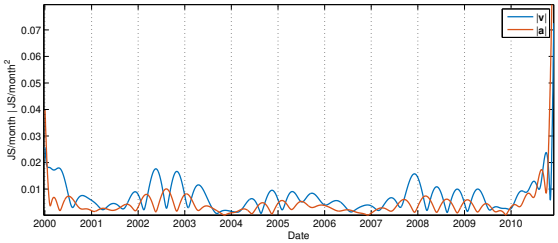
(a) Age IGT plot



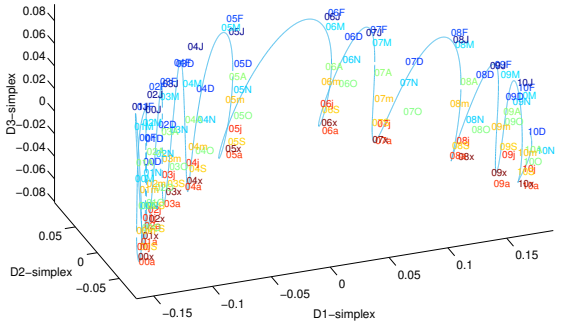
(b) Age kinematic figures



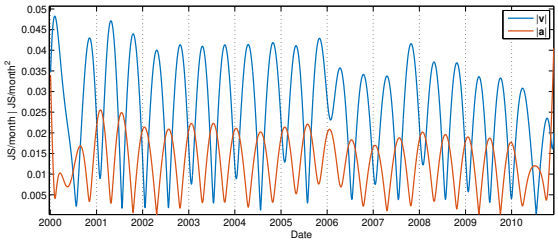
(c) Days of care IGT plot



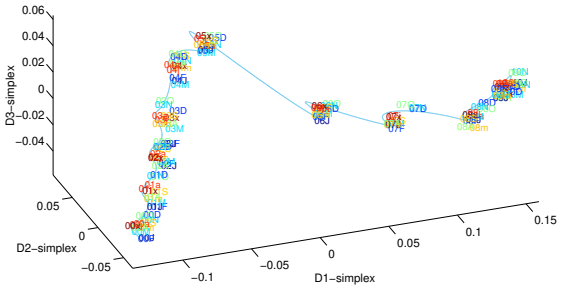
(d) Days of care kinematic figures



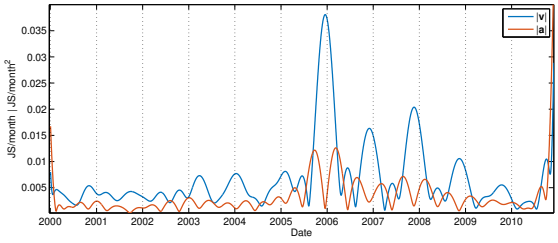
(e) Diagnosis code #1 IGT plot



(f) Diagnosis code #1 kinematic figures



(g) Procedure code #1 IGT plot



(h) Procedure code #1 kinematic figures

Figure 2: Information Geometric Temporal (IGT) plots (left column) and kinematic figures of celerity $\|\mathbf{v}\|$ and acceleration $\|\mathbf{a}\|$ of data changes (right column) for the univariate analysis.

33 months for Days of care, a low-frequency component related to the three waves of change in velocity we can observe in the original kinematic figures (Figure 2(d)), what seems to be also revealed by the corresponding autocorrelation plot (Figure 4(d)). This 33-month component is also found in Age, and seems related to the partition in temporal subgroups observed in IGT plot and three waves of velocity in the kinematic figures (although to a lower degree than in Days of care).

In contrast, neither the Fourier or correlation analysis of individual IGT did reveal any seasonality in Procedure codes, as we may expect given the seasonality of diseases. We can argue that this may be due to the yearly temporal subgroup partition, observed as the isolated *temporal data islands* in the IGT plot (Figure 2(g)), and reflected in its related 12-month periodic changes in velocity (Figure 3(h)). These widespread temporal subgroups are a clear problem to traditional signal processing methods, where prior manual signal pre-processing might help to some degree. However, and as one of the most important results of this work, the APVV method was robust against the extreme sub-grouping and revealed a truly directed seasonality, what could not be found by conventional Fourier or correlation analysis on the signal, as we see next. And additionally, the kinematic analysis of celerity also revealed such a seasonal behaviour in the speeds of change along the temporal subgroups.

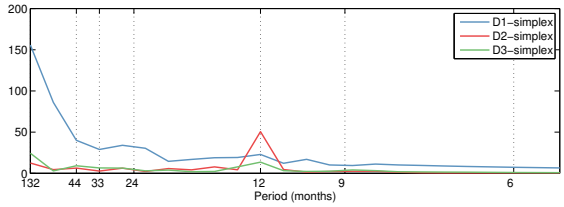
The APVV and APVVmap results discovered new insights about the seasonality of the analysed variables. We already knew the 12-month seasonality at Age, Days of Care and Diagnosis code, however, we are now able to see two novel features describing that seasonality. First, the APVV results (Figure 5, left column) also confirm the 12-month periodic behaviour, but additionally, we now measure if the trajectory is oriented towards the same change direction at the different lags. This is clearly observed for Age, where for $k = 12x : x \in [0, 10]$ ($k = 12$ and its multiples) there is a strong Parallelism, i.e., every 12 months data distributions tend to change towards the same direction. However, for $k = 6 + 12x$ there is an anti-Parallelism, i.e., comparing the direction of change at a specific month with the direction of change at the same month the next year plus 6 months the tendency is opposite. We can observe similar seasonal patterns on APVV for the other variables, however, according to the APVV the anti-Parallelism gets more focused between consecutive years, i.e. for $k = 6$. Besides, APVV plots for Diagnosis code and Procedure code seem to be a composition of two signals. This was tested by performing a Fourier transform on these APVV plots, revealing two frequency peaks, one at 12-month period and the other between 8 and 9 months (Figure SM-7).

Thanks to the APVVmap, we can focus on the specific dates encompassing the Parallelism behind the APVV, although as mentioned in Section 3.2 the APVV counts with lower resolution, so this is taken as a complementary result. We observe that for all the variables the months at which the tendency of change is mostly parallel in a yearly basis (lag of 12) are those at the middle of the year, i.e. May-June, and those at the end of the year, i.e. November-December. For example, the distribution of Diagnosis codes tends to change towards the same direction in May and November, yearly. This is also observed for the Procedure codes, what we were unable to detect based on conventional Fourier and autocorrelation analysis in our study. Additionally, as initially revealed by the APVV plot, we are now able to observe for Diagnosis and Procedure codes the component of change at a period of around 9 months (reflected also at $k \approx 17$ months), with peaks of start-end months around September-June. This approximately 9-month period finding could be related to the pregnancy period, as studied in previous works[5].

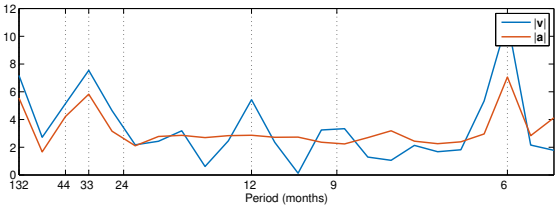
5.2.3 Explanation of temporal variability components

In the previous sections, the findings raised by the estimated data trajectories are related to changes in directions throughout the dimensions of the IGT space. In this section, we relate the dimensions of the IGT embeddings of the analyzed NHDS variables, with the three explanatory factors proposed in the method in Section 3.3. Figure 6 shows the explanatory matrices $\boldsymbol{\mathcal{R}}$ of each variable. These are displayed by means of heatmaps about the resultant correlations. The columns indicate the three first dimensions of the IGT spaces, and the rows the explanatory factors of Date (F1), Seasonality at 12 months (F2) and Temporal subgroups (F3). In this section, we show the results of the six variables.

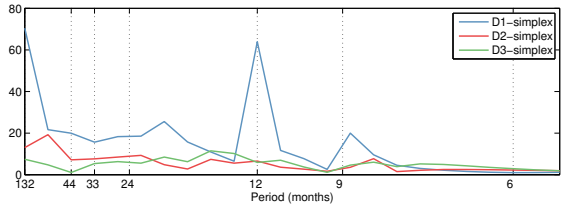
As the clustering method for F3, we used the DBSCAN [45] algorithm. As a compromise between full-automation and accurate results (we could have obtained more accurate clusters by manually setting parameters on each variable), and considering the relationship point-month, the DBSCAN parameters were set as follows. The minimum number of points per cluster was set in three months. The distance threshold ϵ was set as the weighted sum of the median distances from each month to its 12 closest month



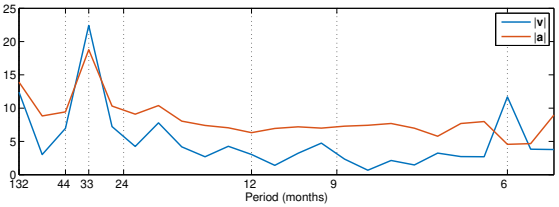
(a) Age IGT components FFT



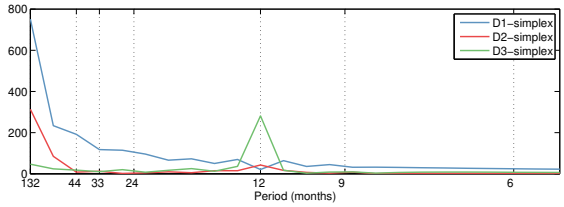
(b) Age kinematic figures FFT



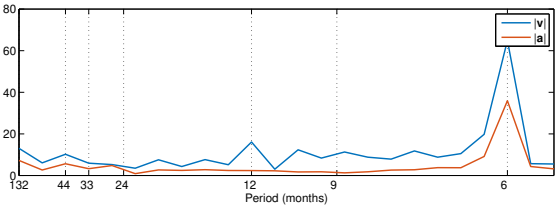
(c) Days of care IGT components FFT



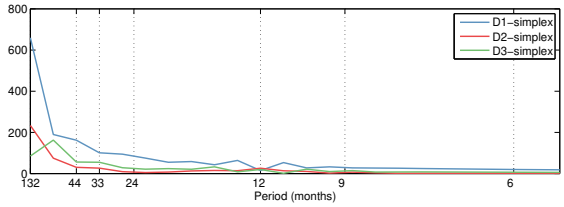
(d) Days of care kinematic figures FFT



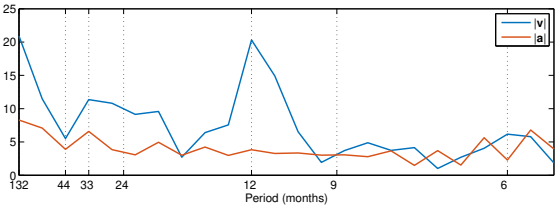
(e) Diagnosis code #1 IGT components FFT



(f) Diagnosis code #1 kinematic figures FFT



(g) Procedure code #1 IGT components FFT

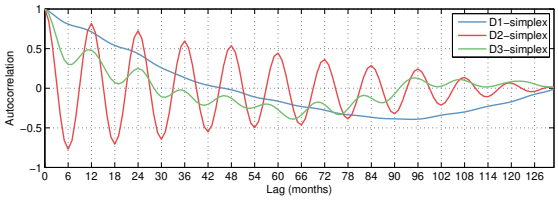


(h) Procedure code #1 kinematic figures FFT

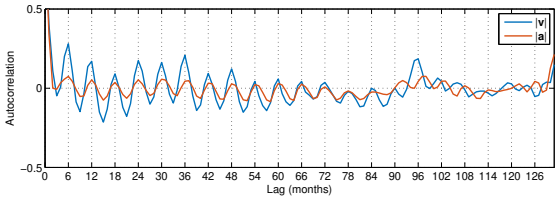
Figure 3: Fourier analysis for individual IGT components (left column) and kinematic figures (right column).

batches. Points not assigned to any cluster were discarded. Temporal subgroups results using this method are provided in the Figure SM-8.

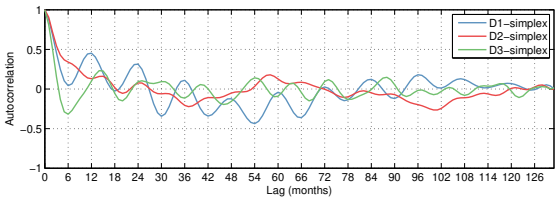
In general, in Figure 6 we observe that the first dimension (D1) concentrates the largest amount of correlation for the different factors. This can be expected since the MDS embedding is made in terms of explained variance. However, we can find some differences among variables. For Age, D1 is explained by the date flow (F1) and the abrupt change in 2008 (F3), as observed in its IGT plot. Seasonality (F2) is also present for Age, mainly in its two first dimensions, according to the levels shown in its autocorrelation plot for $k = 12$ (Figure 4). However, in that figure, we also see that the autocorrelation for D1 is more caused by the signal downward trend at D1 than a seasonal effect, as better explained in the related Fourier transform (Figure 3(c)). In sex, D1 shows a strong negative correlation with F1, what means that time flow is explained by D1 but in the opposite direction on the axis, as compared with the other variables. Temporal subgroups are also correlated to all the dimensions in Sex, despite the main trend in D1 (Figure SM-3, SM-8). Days of care dimensions show a minor correlation in general in comparison to the other variables, as no patterns are clearly observed in its IGT plot. However, for D2 there is some correlation related to the seasonal effect (F2), as found in the previous results. Diagnosis and Procedure codes show a strong correlation of time flow, as their trend on D1 direction, but temporal subgroups are



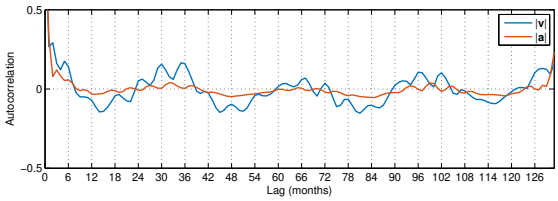
(a) Age IGT components autocorrelation



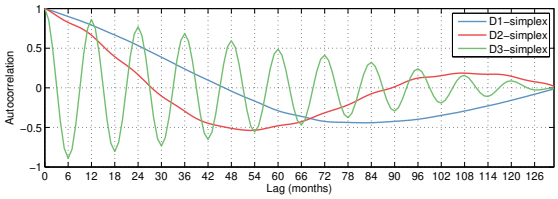
(b) Age kinematic figures autocorrelation



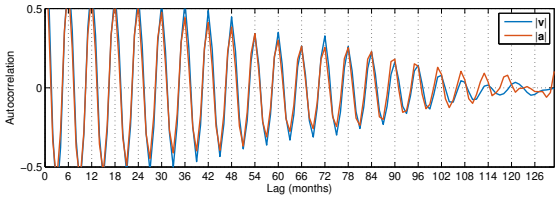
(c) Days of care IGT components autocorrelation



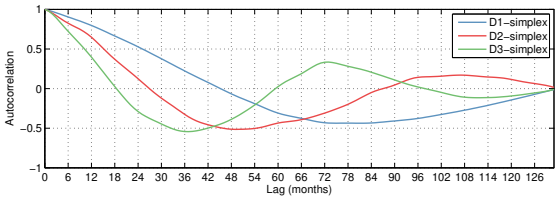
(d) Days of care kinematic figures autocorrelation



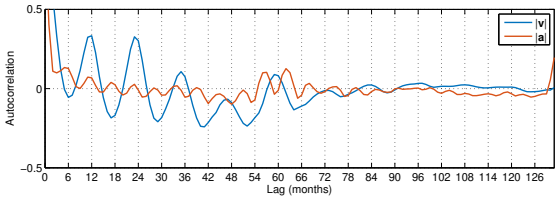
(e) Diagnosis code #1 IGT components autocorrelation



(f) Diagnosis code #1 kinematic figures autocorrelation



(g) Procedure code #1 IGT components autocorrelation



(h) Procedure kinematic figures autocorrelation

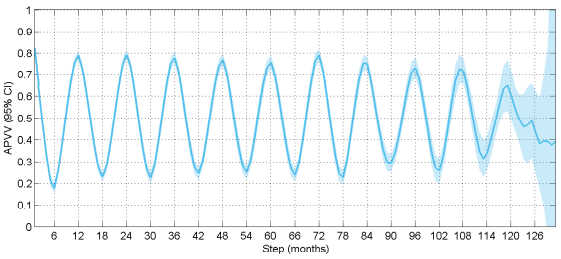
Figure 4: Autocorrelation analysis for individual IGT components (left column) and kinematic figures (right column).

also a major source of variance on D1, as this can be clearly observed in Figure SM-8. Seasonality is also explained in D3 for Diagnosis code, as we clearly found in previous results. Interestingly, time flow (F1) has no correlation with D2 and D3. Procedure code, characterized by the yearly temporal subgroups, show a high correlation for time flow (F1) and these subgroups (F3) on D1, although subgroups are well explained too by the other two dimensions, as observed in the subgroups plot. Seasonality, hard to find visually, is also explained to some degree by D1 and D2. Finally, DRG is also characterized by strong temporal subgroups, as observed in its plots, with a degree of time flow on D2 and D1.

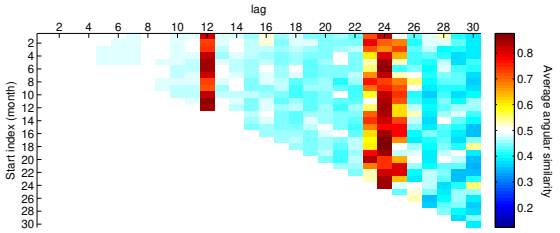
We must recall here that a component of change, i.e. such as F1 to F3, shall not be only explained by a specific dimension. In fact, some components could be explained as a linear combination of several dimensions, such as the seasonality. Therefore, vectors describing an orientation of change on the estimated IGT series, such as the velocity vectors, could be explained by weighting $\boldsymbol{\mathcal{R}}$ by their specific dimensional magnitudes.

6 Discussion

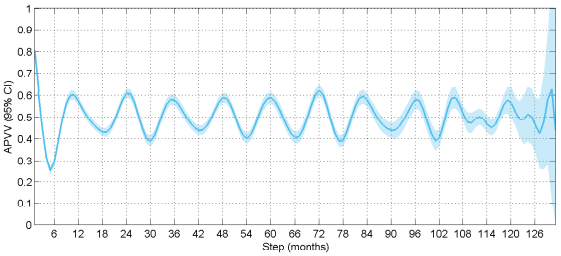
Temporal factors such as changes in protocols, clinical practice, population changes, or data quality issues, can be a potential source of bias when reusing Big Biomedical Data repositories for knowledge discovery.



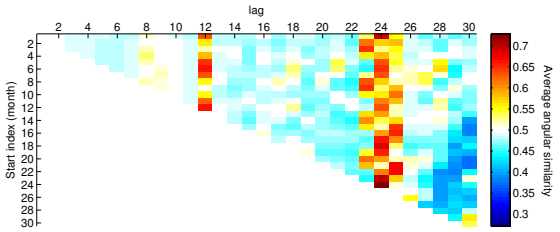
(a) Age APVV



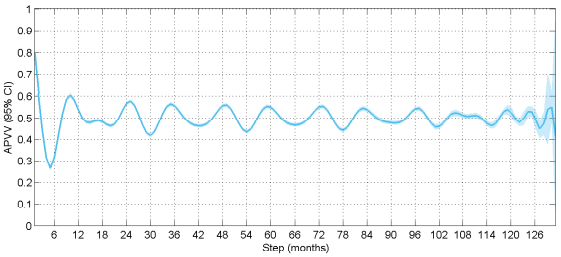
(b) Age APVVmap



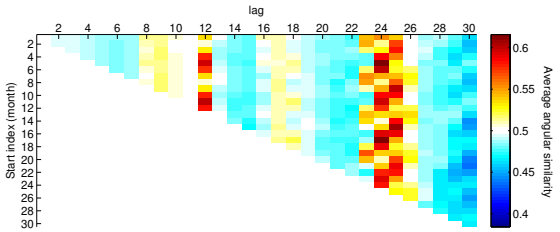
(c) Days of care APVV



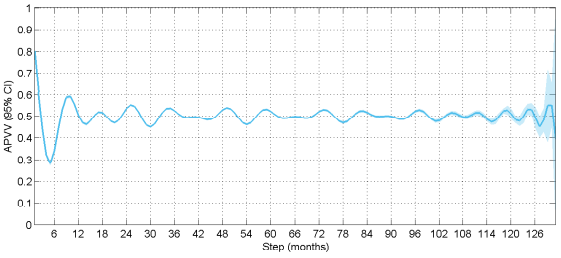
(d) Days of care APVVmap



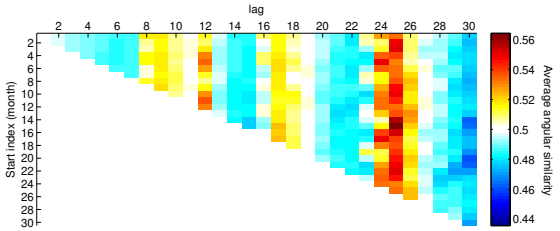
(e) Diagnosis code #1 APVV



(f) Diagnosis code #1 APVVmap



(g) Procedure code #1 APVV



(h) Procedure code #1 APVVmap

Figure 5: Results of the Auto-Parallelism of Velocity Vectors, including APVV plots (left column) and APVVmaps (right column).

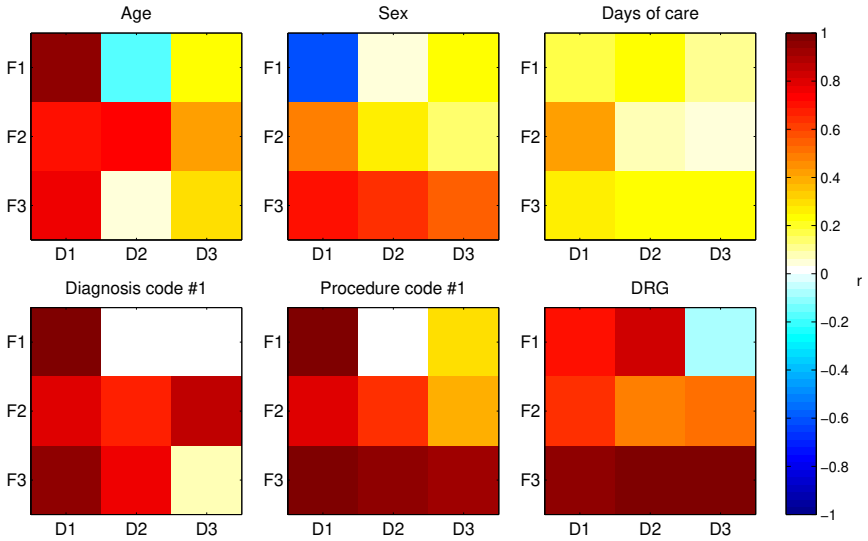


Figure 6: Contribution of different temporal factors to the first three dimensions of the IGT space for the different variables. r: dimension-factor Spearman correlation, F1: Date, F2: Seasonality at 12 months, F3: Temporal subgroups.

This is particularly important in research, and more concretely in statistical modelling. For instance, if building a predictive model for patient 30-day readmission based on NHDS discharge data, those variables showing unstable temporal behaviours, such as the temporal subgroups found in Procedure codes and DRG, or variables with strong seasonal effects, should be considered in a way that such variability in the training set is not affecting the model reliability on further data. This is only an example of why we stress that previous to any data science research, researchers should be aware of what and how is temporal variability of data. This will be even more important in the coming years since Big Biomedical Data repositories will count with data acquired over longer periods of time.

6.1 Significance

Temporal variability of data is traditionally addressed based on conventional statistical and time series methods. Researchers could measure the degree of change on a numerical variable such as Age by means of monitoring its mean over time, e.g., trends could be studied by means of a moving average, while seasonality by means of autocorrelation. However, Age, even within a restricted clinical domain, could stand multi-modal, and summarizing a multimodal distribution on a single statistic such as the mean, or even the mode, entails a loss of information. In the case of categorical variables, researchers could address their temporal variability by monitoring the frequencies of their individual values, likewise as in age. In this case, the loss of information is given by leaving out the rest of categorical values.

In previous works, we proposed addressing the temporal variability of biomedical data by means of embedding time batched statistical distributions (independently of the type of variables and even on multivariate mixed data) in a statistical manifold, leading to the IGT plots and a statistical process control method. However, these mainly addressed visual analytics and required a reinforcement in terms of explaining and interpreting changes.

The adopted solution, as proposed in this paper, led to a kinematic model of biomedical data, and relies firstly on the estimation of a continuous trajectory of data temporal evolution through the IGT plot, by means of FDA. The originally multidimensional and space-embedded (coordinates without explicit meaning) nature of this series, made conventional time series methods such as autocorrelation or Fourier analysis to be insufficient. In consequence, we created the APVV and APVVmat methods, which are multivariate and coordinate-free (robust against rotations of the signal in its space) seasonality methods, which measure the oriented seasonality towards a specific direction of change in the original space. In fact, the APVV and APVVmat can be used in conjunction to Fourier and autocorrelation analysis being applied to the kinematic figures, leading to two types of seasonality: in the change tendency or orientation, and in the degree of change (based on velocity and acceleration).

We specially stress the robustness of the proposed kinematics and APVV methods against extreme temporal sub-grouping, such as found in the variable Procedure code, where the APVV method revealed two directed seasonal periods at 12 and 9 months, what could not be found by conventional Fourier or correlation analysis on the raw individual coordinates (even complex manual signal preprocessing methods would find problems in addressing or removing those temporal subgroups).

The results on the NHDS were aligned with the dataset and population descriptions found in the literature. In addition, the evaluation allowed emphasizing the potential of the proposed methods to provide novel wide-population, and Big Data specialized results. It can be highlighted that the APVV revealed seasonal behaviours in cases where conventional Fourier and autocorrelation analysis were not able, such as in the 12-month and 9-month periods in Diagnosis and Procedure codes, even being robust to a large number of temporal subgroups in the latter. Additionally, the APVVmap allowed detecting the specific start-end months with a major contribution to the seasonality at a specific lag.

Section 2 positioned our work with respect to background work. However, we would like to highlight here some specific relationships to other works. Ours is not the first work where trajectories of medical data are studied, however, most studied trajectories of individual patients, instead of a population-wide trajectory. As an example, some works modelled the disease trajectories of patients with Autism Spectrum Disorder (ASD) under a topic model approach (i.e., assuming an evolution of a "bag of words" describing patient status over time) using graphical models such as Latent Dirichlet Allocation [46, 47], and aimed to ASD subphenotyping. Despite the patient-level approach, further work could be carried out to investigate the use of graphical models in our population-level approach. We would also like to make reference to the work by CC Aggarwal [48], to our knowledge the first proposing a physical model for changes in data distributions. That work introduced the idea of velocity density estimation to estimate the rate at which changes in probability density occur at spatial locations of the distribution support. The main difference with our work is that our model addresses the kinematic figures of the full distribution, i.e., considering population data as a live particle moving over the IGT space over time.

The implications of this work in the reuse of Big Biomedical Data are driven by the amount of information that is provided characterizing temporal variability. This could help to focus on specific time periods to avoid batch effects or detect multi-dimensional seasonal patterns what could entail unexpected drawbacks for data analytics. Kinematic figures of changes could be used in benchmarking data, even real-time acceleration could be used to advance short-term changes. In this regard, the estimated trajectories could also be tracked and used for prediction using FDA or traditional predictive models, although with caution with the predicted confidence intervals. However, we believe that most implications could be for derived work in biomedical data science research. Having measured seasonal effects, we can get to removing these effects on the estimated data distributions to focus on overall trends, e.g., by inverse Fourier transform on each probability bin as a time series. This could also be used to simulate reconstructed raw data without seasonal effects, based on Monte-Carlo sampling from these seasonally-curated distributions. On the other hand, additional work on the relationship of velocity vectors into the IGT space could lead to explaining the exact probability masses establishing the tendency of changes. Also focusing on the raw data, unsupervised learning methods could be applied to the probability heatmaps behind the trajectory estimation (see Figure 1(a)), so that specific trajectories can be derived for isolated patient subgroups. Finally, by means of using purely parametric Information Geometry, the IGT plot coordinates would indeed have a meaning, as the parameters of their projected distributions, what would facilitate the explanation of changes, however, modelling the heterogeneous features of biomedical data into a parametric distribution could entail a complex probability model.

We would also like to highlight a possible implication that this work may have to Public Health research, what would come by switching the units of measurement used in the IGT spaces. Interestingly, if instead of a probabilistic distance, we use an $L1$ (Manhattan or City Block) distance as a pairwise distance between the absolute histograms of time batches (binned counts of individuals within some variable values), such a distance value will be related to the difference in individuals between the histograms. Therefore, the unit of measurement in the derived IGT space will be given as well in individuals (e.g., $individuals/month$). This L1 approach could show interesting analytic properties since the magnitude of motion vectors could be directly translated to patients, and even to costs.

This work could also contribute to other areas of research. We see particularly interesting its application to human and especially to animal tracking [49], where the proposed APVV and APVVmap methods

for trajectory analysis could be applied. These could also provide supporting tools for the understanding and visualization of complex, multivariate seasonal behaviours of live organisms [50].

6.2 Limitations

One limitation of this work is that we did not analyse possible multivariate interactions between the NHDS variables. Although the multivariate analysis already revealed several temporal variability characterizations, their interpretation is not trivial, given the initial dimensionally reduced space. Nevertheless, the interpretation of interactions is complex in most data analysis tasks when dealing with a high number of variables. In this manner, it would be advisable to perform any iterative, possibly automated process from higher to univariate dimensions through which discover the level at which a specific variability finding is explained. Following with the evaluation, some readers would miss an evaluation on artificial, simulated data testing the specific features the methods claim to do. However, we believe that the structured evaluation on the NHDS data has led to a sufficiently thorough validation of all these features.

One additional technical limitation was revealed while analyzing the Days of care variable (Section 5.2.1). Firstly, as a numerical variable, the estimation of its density is sensitive to outliers, what happened with extreme values of it. While setting the number of bins for the distribution estimation, this should be considered to take care of the balance between resolution (not losing detail at high information regions) and coverage (not missing extreme values which, in the end, are real data). And secondly, we found that specific month with extreme distribution (see Figure SM-9), what can be indeed real given to punctual situation, can affect the results of the trajectory estimation, and in consequence to the rest of the analysis. While this should not be considered a problem at all, since it is a real situation, we could take two consequent actions. The first, we could consider an iterative analysis when these situations occur by manually fixing and repeating the analysis. Second, we could attempt to automatically redo an analysis by allowing FDA to smooth out that "noisy" point, or to previously apply a smoothed temporal distribution estimation such as we proposed in a previous work [5].

Finally, we have explained the reasons that to our knowledge explain the variability findings in the NHDS data. However, a deeper analysis from a clinical perspective could support these explanations while further explaining relationships between real-life-variations and data-variations.

7 Conclusion

Big Biomedical Data repositories acquired over long time periods can show multiple evolving and seasonal patterns in data, because of causes such as changes in protocols, clinical practice, population changes, or data quality issues. These temporal variability factors can be a potential cause of bias if not disclosed or properly managed when reusing the data, especially in research and statistical modelling. By establishing a novel paradigm for biomedical data analytics, we proposed a set of methods to analyse the kinematics of Big Biomedical data towards the characterization of data temporal variability for a reliable data reuse.

Applied to the NHDS open repository, including 3,25M hospital discharges over 11 years, the proposed methods revealed several temporal variability patterns. Data are partitioned into several temporal subgroups, mainly related to changes in coding. That was particularly of impact for the Procedure and DRG codes, which show yearly partitioned temporal subgroups. The kinematic figures of data allowed quantifying the degree of change over time as well as to uncover seasonal effects on the velocity of data evolution. The combination of conventional Fourier and autocorrelation analysis with the newly proposed APVV and APVVmap methods allowed describing and quantifying the seasonality of data distribution changes. The APVV allowed measuring oriented seasonal changes on the IGT space. For most variables, their data distributions tended to change to the same direction at a 12-month period, with a peak of change directionality at mid and end of the year. Diagnosis and Procedure codes also included a 9-month periodic component. Three main factors of change were associated with each dimension of the IGT spaces towards the interpretation of the aforementioned directions of change in such space. Kinematics and APVV methods were able to detect seasonal effects on extreme temporal subgrouped data, such as in Procedure code, where Fourier and autocorrelation methods were not able to.

The increasing availability of Big Biomedical Repositories acquired over larger periods of time is opening the way to scientific works analysing the effects and imprints of time in data. We foresee that the

benefits of specialized temporal variability characterization and visualization methods, as those proposed in this paper, will be of utmost importance in the coming decades in Big Biomedical Data research.

What was already known on the topic - Big Biomedical Data repositories acquired over long time periods can show multiple evolving and seasonal patterns in data, because of causes such as changes in protocols, clinical practice, population changes, or data quality issues. These temporal variability factors can be a potential cause of bias if not disclosed or properly managed when reusing the data, especially in research and statistical modelling. - The heterogeneity of Big Biomedical Data, including multivariate and multi-modal categorical and numerical variables, on a huge number of patients, makes difficult to use conventional time series methods, as well as to deliver proper wide-population visual analytics.
What this study added to our knowledge - By analyzing the kinematics of Data through their trajectories on a multidimensional Information Geometric Temporal (IGT) space we can automatically measure and characterize the temporal variability factors of Big Biomedical Data towards a reliable data reuse. - The Auto-Parallelism of Velocity Vectors (APVV) methods allow explaining the (multi-dimensional) oriented seasonality of data and resulted robust against rotations in the original space (coordinate-free) and against temporal subgrouping, where Fourier and autocorrelation analysis failed. - The Automatic Explanation of Temporal Variability Components method helped in determining the major sources of temporal variability for each case and assigning meaning to the coordinates of the IGT spaces. - An evaluation of temporal variability factors of the US National Hospital Discharge Survey (NHDS) open repository, including 3,25M hospital discharges over 11 years, describing its temporal kinematics, temporal subgroups, and oriented seasonality.

Table 2: Summary table

Acknowledgements

This work was supported by UPV grant No. PAID-00-17, and projects DPI2016-80054-R and H2020-SC1-2016-CNECT No. 727560.

References

References

[1] V. Gewin, Data sharing: An open mind on open data, Nature 529 (7584) (2016) 117–119.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, G. Z. Yang, Big data for health, IEEE Journal of Biomedical and Health Informatics 19 (4) (2015) 1193–1208.

[3] F. Bray, D. M. Parkin, Evaluation of data quality in the cancer registry: Principles and methods. part i: Comparability, validity and timeliness, European Journal of Cancer 45 (5) (2009) 747–755.

[4] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, J. F. Steiner, A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research:, Medical Care 50 (2012) S21–S29.

[5] C. Sáez, P. P. Rodrigues, J. Gama, M. Robles, J. M. García-Gómez, Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality, Data Mining and Knowledge Discovery 29 (4) (2015) 950–975.

[6] C. Sáez, O. Zurriaga, J. Pérez-Panadés, I. Melchor, M. Robles, J. M. García-Gómez, Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in spain: a systematic approach to quality control of repositories, Journal of the American Medical Informatics Association 23 (6) (2016) 1085–1095.

[7] C. Sáez, M. Robles, J. M. García-Gómez, Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances, Statistical methods in medical research 26 (1) (2017) 312–336.

[8] D. Schlegel, G. Ficheur, Secondary use of patient data: Review of the literature published in 2016, Yearbook of medical informatics 26 (01) (2017) 68–70.

[9] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, R. A. Irizarry, Tackling the widespread and critical impact of batch effects in high-throughput data, Nature Reviews Genetics 11 (10) (2010) 733.

[10] L. Jacob, J. A. Gagnon-Bartsch, T. P. Speed, Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed, Biostatistics 17 (1) (2015) 16–28.

[11] W. W. B. Goh, W. Wang, L. Wong, Why batch effects matter in omics data, and how to avoid them, Trends in biotechnology 35 (6) (2017) 498–507.

[12] D. Agniel, I. S. Kohane, G. M. Weber, Biases in electronic health record data due to processes within the healthcare system: retrospective observational study, BMJ 361.

[13] S.-I. Amari, H. Nagaoka, Methods of Information Geometry (Translations of Mathematical Monographs), American Mathematical Society, 2007.

[14] K. M. Carter, R. Raich, W. G. Finn, A. O. Hero, Fine: Fisher information non-parametric embedding, arXiv preprint arXiv:0802.2050.

[15] J. O. Ramsay, B. W. Silverman, Functional data analysis, Springer, New York, 2005.

[16] NHDS, National Center for Health Statistics, National Hospital Discharge Survey (NHDS) data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, available at: http://www.cdc.gov/nchs/nhds.htm (2014).

[17] R. Y. Wang, D. M. Strong, Beyond accuracy: what data quality means to data consumers, J. Manage. Inf. Syst. 12 (4) (1996) 5–33.

[18] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang, Aimq: A methodology for information quality assessment, Inf. Manage. 40 (2) (2002) 133–146.

[19] A. F. Karr, A. P. Sanil, D. L. Banks, Data quality: A statistical perspective, Statistical Methodology 3 (2) (2006) 137 – 173.

[20] S. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A. Yeo, A. Talaei-Khoei, Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature, International Journal of Medical Informatics 82 (1) (2013) 10–24.

[21] N. G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J Am Med Inform Assoc 20 (1) (2013) 144–151.

[22] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, et al., A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data, Egems 4 (1).

[23] B. Heinrich, M. Klier, M. Kaiser, A procedure to develop metrics for currency and its application in CRM, Journal of Data and Information Quality 1 (1) (2009) 1–28.

[24] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Computing Surveys 41 (3) (2009) 1–52.

[25] G. Svolba, P. Bauer, Statistical quality control in clinical trials, Controlled clinical trials 20 (6) (1999) 519–530.

[26] P. Brockwell, R. Davis, Time Series: Theory and Methods, Springer Series in Statistics, Springer, 2009.

[27] M. C. S. Wong, X. Q. Lao, K.-F. Ho, W. B. Goggins, S. L. A. Tse, Incidence and mortality of lung cancer: global trends and association with socioeconomic status, Scientific Reports 7 (1) (2017) 14300.

[28] G. E. Box, G. M. Jenkins, Time series analysis: forecasting and control, revised ed, Holden-Day, 1976.

[29] G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.

[30] M. Basseville, I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[31] J. Gama, M. M. Gaber, Learning from Data Streams: Processing Techniques in Sensor Networks, Springer, 2007.

[32] W. A. Shewhart, W. E. Deming, Statistical method from the viewpoint of quality control, Washington, D.C. : Graduate School of the Department of Agriculture, 1939.

[33] H. Mouss, D. Mouss, N. Mouss, L. Sefouhi, Test of page-hinckley, an approach for fault detection in an agro-alimentary production system, in: Control Conference, 2004. 5th Asian, Vol. 2, 2004, pp. 815–818 Vol.2.

[34] T. M. Mitchell, R. Caruana, D. Freitag, J. McDermott, D. Zabowski, Experience with a learning personal assistant, Commun. ACM 37 (7) (1994) 80–91.

[35] J. Gama, P. Medas, G. Castillo, P. Rodrigues, Learning with drift detection, in: A. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004, Vol. 3171 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2004, pp. 286–295.

[36] P. P. Rodrigues, J. Gama, R. Sebastião, Memoryless fading windows in ubiquitous settings, in: In Proceedings of Ubiquitous Data Mining (UDM) Workshop in conjunction with the 19th European Conference on Artificial Intelligence - ECAI 2010, 2010, pp. 27–32.

[37] H. Sørensen, J. Goldsmith, L. M. Sangalli, An introduction with medical applications to functional data analysis, Statistics in medicine 32 (30) (2013) 5222–5240.

[38] E. Parzen, On estimation of a probability density function and mode, The Annals of Mathematical Statistics 33 (3) (1962) 1065–1076.

[39] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information Theory 37 (1991) 145–151.

[40] D. Endres, J. Schindelin, A new metric for probability distributions, IEEE Transactions on Information Theory 49 (7) (2003) 1858–1860.

[41] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79–86.

[42] W. Torgerson, Multidimensional scaling: I. theory and method, Psychometrika 17 (4) (1952) 401–419.

[43] I. Borg, P. J. F. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer, 2010.

[44] U. D. of Health, H. C. F. A. Humans Services, National Center for Health Statistics, International classification of diseases, 9th revision, clinical modification, 6th edition. (2011).

[45] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226–231.

[46] F. Doshi-Velez, Y. Ge, I. Kohane, Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis, Pediatrics 133 (1) (2014) e54–63.

[47] P. Schulam, F. Wigley, S. Saria, Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery (2015).

[48] C. Aggarwal, A framework for diagnosing changes in evolving data streams, in: ACM SIGMOD Conference, 2003, pp. 575–586.

[49] R. Kays, M. C. Crofoot, W. Jetz, M. Wikelski, Terrestrial animal tracking as an eye on life and planet, Science 348 (6240).

[50] R. J. Nelson, L. L. Badura, , B. D. Goldman, Mechanisms of seasonal cycles of behavior, Annual Review of Psychology 41 (1) (1990) 81–108.