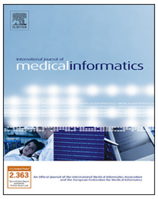
International Journal of Medical Informatics 119 (2018) 109–124

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



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*, Juan M García-Gómez

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

ARTICLE INFO

Keywords:
Temporal stability
Data quality
Time series
Data reuse
Big data
Seasonality
Coordinate-free
Trajectories
Functional data analysis
Statistical manifolds

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 and 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.

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

* Corresponding author.
E-mail addresses: carsaesi@upv.es (C. Sáez), juanmig@upv.es (J.M. García-Gómez).

https://doi.org/10.1016/j.ijmedinf.2018.09.015
Received 17 May 2018; Received in revised form 5 September 2018; Accepted 13 September 2018

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 [3–7], 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 [9–11], 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, multi-modal 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 [17–19]. Several works have reviewed the DQ literature regarding dimensions for the reuse of biomedical data [20–22]. 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