MODELLISTICA E CALCOLO SCIENTIFICO
MODELING AND SCIENTIFIC COMPUTING

MOX-Report No. 21/2021

A General Bi-clustering Algorithm for Hilbert Data: Analysis of the Lombardy Railway Service

Torti, A.; Galvani, M.; Menafoglio, A.; Secchi, P.; Vantini S.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

A General Bi-clustering Algorithm for Hilbert Data: Analysis of the Lombardy Railway Service

Agostino Torti[1,2,*] Marta Galvani[1,*]
Alessandra Menafoglio[1] Piercesare Secchi[1,2]
Simone Vantini[1]

[1]MOX - Department of Mathematics, Politecnico di Milano
[2]Center for Analysis Decisions and Society, Human Technopole, Milano

Abstract

A general and flexible bi-clustering algorithm for the analysis of Hilbert data is presented in the Object Oriented Data Analysis framework. The algorithm, called HC2 (i.e. Hilbert Cheng and Church), is a non-parametric method to bi-cluster Hilbert data indexed in a matrix structure. The Cheng and Church approach is here extended to the general case of data embedded in a Hilbert space and then applied to the analysis of the regional railway service in the Lombardy region with the aim of identifying recurrent patterns in the passengers' daily access to trains and/or stations. The analysed data, modelled as multivariate functional data and time series, allows to measure both overcrowding and travel demand, providing useful insights to best handle the service.

Keywords: Bi-clustering, Clustering, Functional Data, Mobility, Railway Network, Crowding

1 Introduction

Due to urbanization and globalization, in recent years, the demand for transportation has increased like never before. Many benefits to individuals, communities and the local economies are brought by public transport systems. To increase public transport usage, public transport systems need to become more competitive improving the service quality (1). For instance, since overcrowding is one of the major cause of passengers dissatisfaction (2), a first step to improve service quality, is to measure the quality of the service in order to identify the potential weaknesses of a public transport system. The aim of this work is to develop a strategy for profiling the mobility flows and identifying situations of crowding in a railway system.

Since mobility data are continuously collected over time and space, they are often studied as time series or functions, see for example (3), (4), (5) and (6). More in general, this increasing availability of complex and high-dimensional data has motivated a fast and extensive growth of Object Oriented Data Analysis (OODA, e.g., (7)), a branch of statistics which seeks to observe each "object datum" - in our case, time series and functions - as a realization of a random element in a finite or infinite-dimensional space. To extract mobility patterns and give a synthetic view of the information contained in this type of data, clustering approaches are usually employed in the literature (e.g., (8), (9), (10), (11)). With such approaches the whole set of statistical units of interest (e.g. stations) can be stratified in smaller subsets according to specific mobility flows. Nevertheless, the interpretation of these clusters can be difficult due to the long period of observation; hence, cutting this period into small units, e.g. days, and arrange the data in a data matrix, where rows are the statistical units of interest and columns are time units, can help in understanding the phenomenon. In this framework, bi-clustering techniques, first introduced by (12) for expression data, allow to simultaneously group the rows and the columns of a data matrix when data are intrinsically ordered in a matrix structure. Classical bi-clustering methods have been developed to deal with data matrices whose entries are scalar. When the entry of each cell is a data object more complex than a scalar, new approaches within the framework of OODA should be developed. When dealing with time series or functional data, there are just few works dedicated to the bi-clustering in data matrix where each cell contains a single curve: (13) and (14) develop a parametric bi-clustering technique, based on the functional latent block model ((15)), to co-cluster different electricity consumption curves on different days, while (16) develop a non parametric bi-clustering technique, based on the extension of the Cheng and Church algorithm (12), to co-cluster different bike station usage profiles on different days. Alternatively, (17) develop an algorithm that permits to find subsets of functions that exhibit similar behaviour across the same continuous subsets of the domain without assuming any matrix structure in the data. In all these works the analysed objects are functions with one-dimensional domain and codomain.

In this work, to characterise mobility flows and identify problems of overcrowding, we observe the different stations and the scheduled trains of a railway infrastructure on a period of nine days. Moreover, for each statistical unit we consider and analyse multiple aspects of the phenomenon, measuring both overcrowding and travel demand, thus obtaining multivariate functional data and multivariate time series. To consider this set of information, we need to employ a strategy able to bi-cluster a data matrix where in each cell an object, possibly belonging to a multidimensional space, is contained. To the best of our knowledge, there are no works in the literature concerned with the problem of bi-clustering generic object data. Hence, we present a bi-clustering method called the HC2 (i.e, Hilbert Cheng and Church) that can be applied to the analysis of object data for which a meaningful Hilbert space structure can be identified. From a methodological point of view, we first extend the concept of ideal bi-cluster in the framework of Hilbert data, then, we define a suitable index (i.e., H-score) to measure the discrepancy of a generic bicluster (i.e., a selection of rows and columns) to its ideal counterpart, and finally we propose an algorithm seeking for biclusters in the data matrix with low H-score. From an application point of view, we carry out two complementary analyses, focusing, respectively, on the analysis of stations and trains. First, we study the passengers' departures and arrivals at each day-hour for each station along different days. This allows us to identify subsets of stations that in specific days show similar patterns of departures and arrivals along the day point out station-day pairs that could be homogeneously managed by the railway service provider; similarly, for a given line we study also the passengers' boarding, deboarding, and occupancy of each scheduled train along its journey in different days, to identify groups of trains that in specific days show a similar usage profile across the line stations. These two different approaches allow us to have a complete view of the system, identifying eventual issues in specific stations and days, and specific scheduled trains in the different days.

The manuscript is structured as follows: in Section 2 the analysed railway service and infrastructure is presented, coupled with a description of the collected information and the data elaboration procedure. The HC2 algorithm is then presented in Section 3. To show the potential of the method, a simulation study is shown in Section 4. Data are analysed in Section 5. Conclusions are presented in Section 6.

2 Analysis of railway travel demand in Lombardy

With an area of 23.844 square kilometres, a population of 10 million (more than one-sixth of Italy's population) and a fifth of Italy's GDP produced in the region, Lombardy is considered the most populous, richest and most productive region in the country, and one of the EU engines In 2019, about 17 million trips per day, considering different modes of transport, were counted across the region, making Lombardy also the region of Italy with the highest number of daily trips (`https://www.dati.lombardia.it`). The major company responsible for the management of regional passenger train operations in Lombardy is Trenord. Considering only the regional transport, the company operates almost 45 regional lines, 15 suburban lines and the Malpensa Airport express line. Trenord operates almost 2200 rides per day, serving almost 820.000 people (during working days) on the 1.920 km long regional railway network. In Figure 1 the whole railway infrastructure is represented showing stations and rail location. The infrastructure covers areas different in terms of functionality and urbanization, connecting the Milan metropolitan area (the largest in Italy and the third most populated functional urban area in the EU) and the others main cities of the region to rural and mountain Alpine areas spread around.

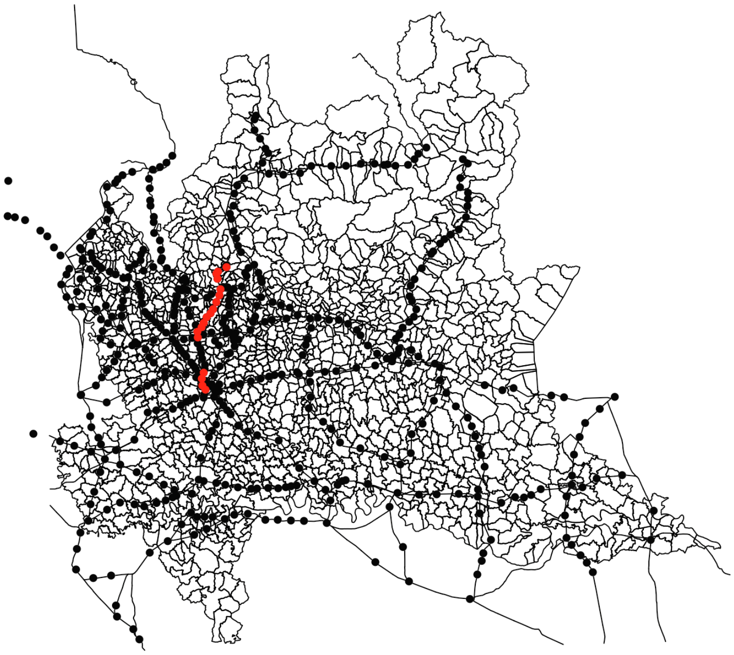

Figure 1: Stations and rail locations in Lombardy. Stations of the Cadorna - Asso line are highlighted in red.

2.1 People counter data

This work focuses on people counter data. In details, for each train in each station, the number of boarding and dropping passengers is measured and provided with extra meta-information on the train, such as the maximum capacity. We then obtain, for each station and train, information on the load factor, i.e. how much of a train passenger carrying capacity is used, and the travel demand, observing boarded and dropped passengers at each station. Data are available for a period of 9 consecutive days in November 2019 (five working days and four weekend days). The analysis here presented focuses on a specific train service: the Milano Cadorna - Asso line (Figure 1). The Milano Cadorna - Asso line is a railway service with 19 stations from the Milan city center (i.e, Milano Cadorna station) to the first Southern slopes of the Alps (i.e., Asso station). This service passes through the provinces of Milan, Monza, and Como with a 50 km route of about 78 minutes duration in each direction. The number and the schedule of the trains, in each direction, change according to the day of the week, going from a minimum of 11 to a maximum of 20 trains per day. In the analysed period nearly 125.000 passengers used on this train service, with a daily average of almost 21.000 passengers during working days and 3.000 during weekend days.

It is convenient to model a train service as a directed network where the nodes are the ordered stations along the line, $S = \{s_1, ..., s_n\}$, and the edges are the railway segments in between two consecutive stations, (s_k, s_{k+1}), with $s_k,\ s_{k+1} \in S$. We then define Q as the set of days in the analysed period and Z as the set of scheduled trains during the analysed period travelling along the line. Fixed a day $q \in Q$, for each train $z \in Z$, we know at each station $s_k \in S$ the number of boarded passengers $b^z_{s_k}$ and the number of dropped passengers $d^z_{s_k}$. From these data, we can evaluate for each edge of the directed network the number of passengers travelling on a specific train z between two stations s_k and s_{k+1} belonging to S as:

$$flow^z_{s_k,s_{k+1}} = \sum_{i=1}^{k}(b^z_{s_i} - d^z_{s_i}).$$

A scheme is represented in Figure 2. For each train we also know the train capacity, C^z, i.e. the maximum number of passengers allowed on a train. Using these data, we can evaluate the load factor for a specific edge as the passengers flow on the edge divided by the train capacity:

$$LF^z_{s_k,s_{k+1}} = flow^z_{s_k,s_{k+1}}/C^z.$$

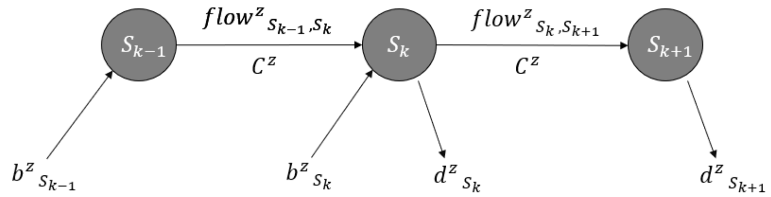

Figure 2: Representation of the available data

2.2 Framing the station data as multivariate functional data

When observing a station along a day, we would like to analyse along time the load factor of the trains arriving and departing from the station and the usage of the station in terms of boarded and dropped passengers. Note that we are interesting in the analysis of this phenomenon on a hourly temporal scale and not minute by minute, even thought we potentially have the data each time a train is entering in a station; indeed, these latter data can be too oscillating and noisy at the risk of loosing the information we are interested in. To this end, for each station $s_k \in S$, for each day and each hour of the day $h \in \{$00:00-00:59, ... , 23:00-23:59$\}$, we observe the load factor of incoming and outcoming trains, $LF_{s_{k-1},s_k}(h)$ and $LF_{s_k,s_{k+1}}(h)$ respectively, as the total hourly incoming and outcoming flows over the total capacity considering all the trains crossing the station in the considered hour. In addition, we also evaluate the boarded/dropped $Balance_{s_k}(h)$, for each hour h of the day in the station $s_k \in S$, as the difference between the total number of boarded and dropped passengers over the maximum between the total boarded and the total dropped passengers. Formulas for these quantities are:

- $LF_{s_{k-1},s_k}(h) = \frac{\sum_{z \in Z_{s_k}(h)} flow^z_{s_{k-1},s_k}}{\sum_{z \in Z_{s_k}(h)} C^z}$;

- $LF_{s_k,s_{k+1}}(h) = \frac{\sum_{z \in Z_{s_k}(h)} flow^z_{s_k,s_{k+1}}}{\sum_{z \in Z_{s_k}(h)} C^z}$;

- $Balance_{s_k}(h) =$

$$= \frac{\sum_{z \in Z_{s_k}(h)} (b^z_{s_k} - d^t_{s_k})}{max(\sum_{z \in Z_{s_k}(h)} b^z_{s_k}; \sum_{z \in Z_{s_k}(h)} d^z_{s_k})} =$$

$$= \begin{cases} \frac{\sum_{z \in Z_{s_k}(h)} (b^z_{s_k} - d^t_{s_k})}{\sum_{z \in Z_{s_k}(h)} b^z_{s_k}}, & \text{if } \sum_{z \in Z_{s_k}(h)} b^z_{s_k} > \sum_{z \in Z_{s_k}(h)} d^z_{s_k}; \\ 0, & \text{if } \sum_{z \in Z_{s_k}(h)} b^z_{s_k} = \sum_{z \in Z_{s_k}(h)} d^z_{s_k}; \\ \frac{\sum_{z \in Z_{s_k}(h)} (b^z_{s_k} - d^t_{s_k})}{\sum_{z \in Z_{s_k}(h)} d^z_{s_k}}, & \text{if } \sum_{z \in Z_{s_k}(h)} b^z_{s_k} < \sum_{z \in Z_{s_k}(h)} d^z_{s_k}; \end{cases}$$

where $Z_{s_k}(h)$ is the set of trains travelling through the station s_k in the hour h. Note that, for simplicity of notation, in the previous formulas we do not include the observed day as index of these quantities, even if each quantity

is differently defined for each station and each day. These sequence of values along time are then transformed into functional data. For each day and each station, the functional data are obtained by smoothing the 24 hourly values: smoothing is performed by means of a Local Weighted Polynomial Regression with a tri-cube kernel function, a bandwidth equal to 0.75 hours and a first order local polynomial (see, for instance, (18)). In this way, for each station $s_k \in S$ in each day $q \in Q$, we obtain a three dimensional functional datum $\boldsymbol{F_{s_k,q}} = (LF_{s_{k-1},s_k}(\cdot), LF_{s_k,s_{k+1}}(\cdot), Balance_{s_k}(\cdot))$, defined on the daily hourly domain $[0, 24]$. The first two components report information about the load factor of the trains arriving and departing from the station s_k along the day q, while the third component is a balance function which represents the interaction between travellers and the station s_k along the day q; indeed, if this quantity is equal to plus one, passengers are only boarding at that station and trains are filling up; if it is equal to zero, the number of boarding and dropping passengers are equal; if it is equal to minus one, passengers are only dropping at that station and trains are emptying out.
These data are arranged in a matrix structure where each row represents a station $s_k \in S$ and each column represents a day $q \in Q$, see Figure 3 (left).

2.3 Framing the train data as multivariate time series

When observing the trains in the different days, we would like to analyse their load factor and study how passengers board and drop along the line in each station. To this end, for every train and day, we consider the *load factor* of the train when arriving at a station, and the proportion of boarded and dropped passengers with respect to the total of passengers carried by the train along its journey. Fixed a day, these quantities are evaluated for each train z each time the train is crossing a stations $s_k \in S$ as:

- $LF^z(s_k) = \frac{flow^z_{s_{k-1},s_k}}{C^z}$;
- $Boarded - proportion^z(s_k) = \frac{b^z_{s_k}}{\sum_{s_k \in S} b^z_{s_k}}$;
- $Dropped - proportion^z(s_k) = \frac{d^z_{s_k}}{\sum_{s_k \in S} d^z_{s_k}}$;

Note that, coherently with the previous section and for simplicity of notation, we do not include the observed day as index of these quantities, even if each quantity is differently defined for each train and each day. In this way, for each train $z \in Z$ and each day $q \in Q$, we obtain a multivariate time series $\boldsymbol{F_{z,q}} = (LF^z(\cdot), Boarded - proportion^z(\cdot), Dropped - proportion^z(\cdot))$ composed from three longitudinal vectors defined for each stop of the line, each vector with number of components equal to the number of stations along the line. The first vector reports information about the load factor of the trains along the line, to highlight overcrowding events. The second and the third vectors represent the interaction between the travellers on the train and the line, indicating the quotas of boarded and dropped for each station along the line. Obviously, for a given train and a day, the sum of the values of $Boarded - proportion$ along the stations sum to one, and similarly for $Dropped - proportion$.

These data are arranged in a matrix structure where each row represents a station $z \in Z$ and each column represents a day $q \in Q$, see Figure 3 (right).

2.4 Analysed data for bi-clustering

As explained above, given a line and a direction along this line, the analysed data are arranged in two data matrices, see Figure 3. Here, in the first case, rows are stations and columns are days, while in the second, rows are trains and columns are days. These two matrices explore different aspects of the system, allowing the study of the line in its entirety. In both cases, we will look for a bi-clustering of the matrices, aiming at highlighting blocks of stations (or trains) behaving in a similar way over different days. Notice that in both cases the entries of the data matrix are objects taking values in a multidimensional space; multivariate functional data when the focus is on stations and multivariate time series when the focus is on trains. In the next section we present our methodological proposal for the bi-clustering of these types of complex object data.

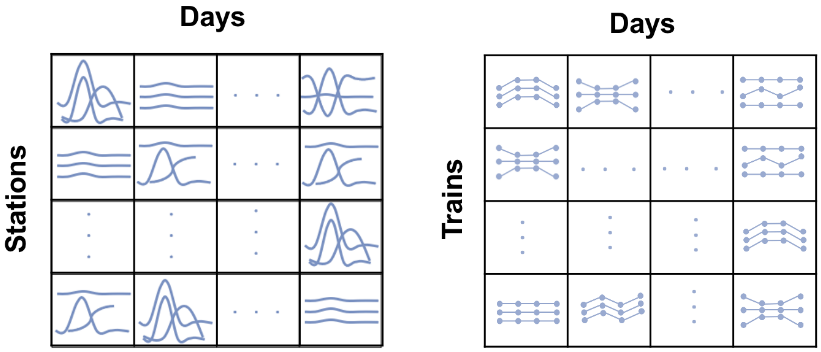

Figure 3: Illustrative visualisation of the obtained data matrices when considering stations along days (left) and trains along days (right)

3 Hilbert Cheng and Church Algorithm (HC2)

To the best of our knowledge, bi-clustering has not yet been considered within the general framework of OODA. A few papers try to bi-cluster data matrices whose entries are functional data, a special case of object data popularized by the seminal work of Ramsay and Silverman (19): (13), (14) and (16). These works proposed different bi-clustering methods extending classical models for multivariate data, in particular (13) and (14) extend the classical Latent Block Model ((15)) to the FunLBM algorithm for functional data, while (16) does the same with Cheng and Church algorithm method, thus generating the FunnCC bi-clustering method. The main difference among these methods is that the FunLBM is a model-based procedure, assuming the existence of a latent-block structure in the data-matrix (which is modeled as a Gaussian mixture distribution on the basis coefficients of the functional data), thus obtaining a semi-parametric method, while FunCC is fully non-parametric. In this work, due to the nature of the analysed data, we rely on the FunCC method as it does not make any prior assumption on the distribution of the data. In addition, through FunCC obtains a non-exhaustive bi-cluster of the data, thus allowing, more realistically, some elements not to belong to any bi-cluster.

3.1 Definition of Bi-cluster for Hilbert Data

Let $\boldsymbol{\mathcal{H}}$ be a Hilbert space with norm $\|\cdot\|_{\boldsymbol{\mathcal{H}}}$ and inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{\mathcal{H}}}$ and $\boldsymbol{\mathcal{X}}$ be a random object belonging to $\boldsymbol{\mathcal{H}}$. Suppose a sample of n x m objects $\boldsymbol{\mathcal{X}_{ij}} \in \boldsymbol{\mathcal{H}}$ is available, arranged in a matrix A composed by n rows and m columns, i.e. each element of the matrix A is an object datum $\boldsymbol{\mathcal{X}_{ij}}$ with $i \in \{1, ..., n\}$ and $j \in \{1, ..., m\}$.

Given a sub-matrix $B(I, J) \subset A$, whose elements are the objects $\boldsymbol{\mathcal{X}_{ij}}$ with $i \in I$ and $j \in J$, we can evaluate the average value $\boldsymbol{\mu}$ of the sub-matrix and the rows/columns components, respectively $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_j}$ as:

$$\boldsymbol{\mu} = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} \boldsymbol{\mathcal{X}_{ij}} \tag{1}$$

$$\boldsymbol{\alpha_i} = \frac{1}{|J|} \sum_{j \in J} \boldsymbol{\mathcal{X}_{ij}} - \boldsymbol{\mu} \tag{2}$$

$$\boldsymbol{\beta_j} = \frac{1}{|I|} \sum_{i \in I} \boldsymbol{\mathcal{X}_{ij}} - \boldsymbol{\mu} \tag{3}$$

Notice that $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_j}$ are objects belonging to $\boldsymbol{\mathcal{H}}$ and representing the rows/columns components, i.e. the residues of respectively rows and columns with respect to the mean value $\boldsymbol{\mu}$ of the sub-matrix.

Definition 3.1 *Given a data matrix A, an ideal bi-cluster is a sub-matrix $B(I, J) \subset A$, s.t. each element $\boldsymbol{\mathcal{X}_{ij}}$ with $i \in I$ and $j \in J$ can be expressed as:*

$$\boldsymbol{\mathcal{X}_{ij}} = \boldsymbol{\mu} + a\boldsymbol{\alpha_i} + b\boldsymbol{\beta_j} \quad, \forall i \in I \;, \; \forall j \in J$$

with $(a, b) \in \{0, 1\}^2$ and $\boldsymbol{\mu}$, $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_j}$ respectively the average value of the $B(I, J)$ and the rows/columns components, defined as in (1)*,* (2) *and* (3)*.*

Starting from Definition 3.1 it is possible to obtain different kinds of ideal bi-clusters, associated to different application contexts, by differently considering a and b. The complete form of ideal bi-cluster, as defined in Definition 3.1, is obtained with $(a, b) = (1, 1)$; considering, instead, $(a, b) = (0, 0)$ in the Definition 3.1, only the average value in the bi-cluster and not the rows/columns components are considered, hence the ideal bi-cluster is composed by a group of objects $\boldsymbol{\mathcal{X}_{ij}}$ all equal to the average value $\boldsymbol{\mu}$ of the bi-cluster. Considering, instead, $(a,b) = (1,0)$ or $(a,b) = (0,1)$, other ideal bi-clusters can be obtained, discovering groups of objects that exhibit coherent variations on the rows or on the columns of the data matrix. This choice is case study specific and depends on the problem at hand.

In practice, we want to find sub-matrices $B(I,J)$ as similar as possible to an ideal bi-cluster, i.e. sub-matrices $B(I,J)$ which minimize a specific objective function. Hence, a specific H-score, which measures the deviation of the selected elements from an ideal bi-cluster, has to be defined. In our case, we define the H-score of a sub-matrix $B(I,J)$ as follow:

Definition 3.2 *Let A be a data matrix and $B(I,J) \subset A$ be a sub-matrix of objects $\boldsymbol{\mathcal{X}_{ij}}$ with $i \in I$ and $j \in J$. The H-score of $B(I,J)$ is defined as:*

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} \left\| \boldsymbol{\mathcal{X}_{ij}} - \boldsymbol{\mathcal{X}^0_{ij}} \right\|^2_{\boldsymbol{\mathcal{H}}}$$

with $\boldsymbol{\mathcal{X}^0_{ij}} = \boldsymbol{\mu} + a\boldsymbol{\alpha_i} + b\boldsymbol{\beta_j}$ being the template object of the sub-matrix $B(I,J)$ with $(a,b) \in \{0,1\}^2$ and $\boldsymbol{\mu}$, $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_j}$ evaluated as in (1), (2) and (3) respectively, using the objects $\boldsymbol{\mathcal{X}_{ij}} \in B(I,J)$.

The introduced H-score evaluates the weighted mean squared residual obtained when representing each element with the estimated template $\boldsymbol{\mathcal{X}^0_{ij}}$ of the sub-matrix to which the object is assigned to.

Note that the definition of ideal bi-cluster and H-score given in (16) are a particular case of the ones given above when the objects of interest are functional data embedded in the Hilbert space L^2.

3.2 The HC2 Algorithm

Given the definitions of ideal bi-cluster and H-score of a sub-matrix, to identify a set of bi-clusters in the matrix structure A, we rely on the same procedure used in (16). The Functional Cheng and Church algorithm (introduced in (16)) extends the original Cheng and Church bi-clustering algorithm (12) to the functional framework by defining a deterministic and non parametric procedure. In this work, we generalize and extend this algorithm for dealing with object data embedded in a Hilbert space, but we refer to (16) for the details, without insisting on the obvious translations to the present more general setting.

In details, the algorithm starts by considering the whole data matrix and proceeds iteratively by removing and adding rows or columns to find the bi-cluster with highest dimensions, in terms of number of rows and columns, and an H-score lower then a given threshold δ. To this end, at each step not only the H-score is estimated but also the row and column scores, necessary to add and remove rows or columns from the considered sub-matrix. In the framework of Hilbert Data, we evaluate the row and the column scores in a sub-matrix $B(I,J)$ as:

$$d_{iJ} = \frac{1}{|J|}\sum_{j \in J} \left\| \boldsymbol{\mathcal{X}_{ij}} - \boldsymbol{\mathcal{X}^0_{ij}} \right\|^2_{\boldsymbol{\mathcal{H}}} \quad \forall i \in I$$

$$d_{Ij} = \frac{1}{|I|}\sum_{i \in I} \left\| \boldsymbol{\mathcal{X}_{ij}} - \boldsymbol{\mathcal{X}^0_{ij}} \right\|^2_{\boldsymbol{\mathcal{H}}} \quad \forall j \in J$$

In practice, the algorithm operates by performing three different iterative procedures: the Multiple Node Deletion phase, the single Node Deletion Phase and the Node Addition phase. In the Multiple Node Deletion phase the algorithm tries to remove at the same time groups of rows or columns with scores bigger than the H-score scaled by a parameter $\theta \geq 1$. In the Single Node Deletion phase the row or the column with the largest score is removed and the H-score of the new obtained matrix is updated; this phase is iteratively repeated until the H-score is lower than a threshold δ. Finally in the Node Addition phase the algorithm tries to add removed rows or columns in order to make the bi-cluster as big as possible without increasing the H-score.

As in the original FunCC algorithm, there are two important parameters, δ and θ, that need to be set before running the algorithm. The parameter δ influences the final number of bi-clusters; when δ is too high a single bi-cluster containing the whole data matrix is obtained, while when is too low a large number of bi-clusters is obtained. If the value of δ is very small, it might even be impossible to find bi-clusters with H-score lower than the thereshold. The parameter θ influences the algorithm speed: when too high only the slower single node deletion step is performed, as it makes impossible to pass through the multiple node deletion step, while when too low the algorithm follows a faster but a too raw procedure, as a high number of rows and columns would be removed in the multiple node deletion step. To tune these parameters the same procedure illustrated in (16) is performed: for the

δ parameter a sensitivity analysis on the number of obtained bi-clusters and the number of not assigned observations is performed, while θ is selected as big as possible, while still allowing for a reasonable computational time.

3.3 The implementation of the method when data objects are multivariate

Notice that, the method presented in the previous section can be applied also in the multivariate case, e.g. dealing with multivariate functional data or multivariate time series. Consider the Hilbert space $\boldsymbol{\mathcal{H}} = \mathcal{H}_1 \times \mathcal{H}_2 \times ... \times \mathcal{H}_P$, with $P \geq 1$, where for $p \in, ..., P\}$, $\mathcal{H}_p$ is a Hilbert space with norm $\|\cdot\|_{\mathcal{H}_p}$ and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_p}$. In this framework, we consider a sample of n x m random objects $\boldsymbol{\mathcal{X}_{ij}} = (\mathcal{X}_{ij}^1, ..., \mathcal{X}_{ij}^P) \in \mathcal{H}_1 \times \mathcal{H}_2 \times ... \times \mathcal{H}_P$, arranged in a matrix A, where $\mathcal{X}_{ij}^p \in \mathcal{H}_p$ for each $p \in \{1, ..., P\}$. In this case, when considering a sub-matrix $B(I, J) \subset A$, we evaluate the H-score and the rows/columns score employing the following norm in $\mathcal{H}_1 \times \mathcal{H}_2 \times ... \times \mathcal{H}_p$:

$$\left\|\boldsymbol{\mathcal{X}_{ij}} - \boldsymbol{\mathcal{X}_{ij}^0}\right\|^2_{\mathcal{H}_1 \times \mathcal{H}_2 \times ... \times \mathcal{H}_p} = \frac{1}{|P|} \sum_{p=1}^{P} w_p \left\|\mathcal{X}_{ij}^p - \mathcal{X}_{ij}^{0,p}\right\|^2_{\mathcal{H}_p} \tag{4}$$

with

$$w_1, ..., w_p \geq 0 \ , \ \sum_{p=1}^{P} w_p = 1$$

and

$$\boldsymbol{\mathcal{X}_{ij}^0} = (\mathcal{X}_{ij}^{0,1}, ..., \mathcal{X}_{ij}^{0,P})$$

being the template object of the sub-matrix $B(I, J)$ with $\mathcal{X}_{ij}^{0,p} = \mu^p + a\alpha_i^p + b\beta_j^p$ for $p \in \{1, ..., P\}$. The choice of weights w_p can be driven by the problem at hand, i.e., when objects in the different dimensions have different units of measure to make the different norm values $\|\cdot\|_{\mathcal{H}_p}$ comparable. Moreover, the weights w_p make also possible to take into account prior knowledge about the data by giving different weights to each component of the object data, guarantying a great adaptivity to the real problem under study.

4 Simulation study

To show the potential of the HC2 model, a data matrix of object data is simulated. The case of multivariate functional data is considered, since in the case study we deal with this kind of data. Specifically, each object is a bi-dimensional functional datum, i.e. for each statistical unit two different functions are observed. A data matrix A of dimensions 20 x 10 with three bi-clusters is obtained as follow:

$$\boldsymbol{a_{ij}} = \begin{cases} \{X^1(t) + \varepsilon^1_{ij}, Y^2(t) + \varepsilon^2_{ij}\} & \forall \ (i,j) \in [1:10, 1:5] \\ \{X^1(t) + \varepsilon^1_{ij}, Y^1(t) + \varepsilon^2_{ij}\} & \forall \ (i,j) \in [1:10, 6:10] \\ \{X^2(t) + \varepsilon^1_{ij}, Y^2(t) + \varepsilon^2_{ij}\} & \forall \ (i,j) \in [11:20, 1:10] \end{cases}$$

where $X^1(t) = (t-2.5)^3/2$, $X^2(t) = (t-2.5)^4/3 - 9$, $Y^1(t) = [-(t^4 - t^3 - 19t^2 - 11t) - 100]/10$ and $Y^2(t) = 6cos(3t)$, with $t \in [0,5]$. The errors ε^p_{ij} with $p \in \{1,2\}$ are from a Gaussian process with zero mean and $E(\varepsilon^p_{ij}(t)\varepsilon^p_{ij}(s)) = e^{-(|t-s|)}$. All other elements in A are i.i.d. noisy data such that $\boldsymbol{a_{ij}} = \{5\varepsilon^1_{ij}, 5\varepsilon^2_{ij}\}$. Resulting simulated data are shown in Figure 4.

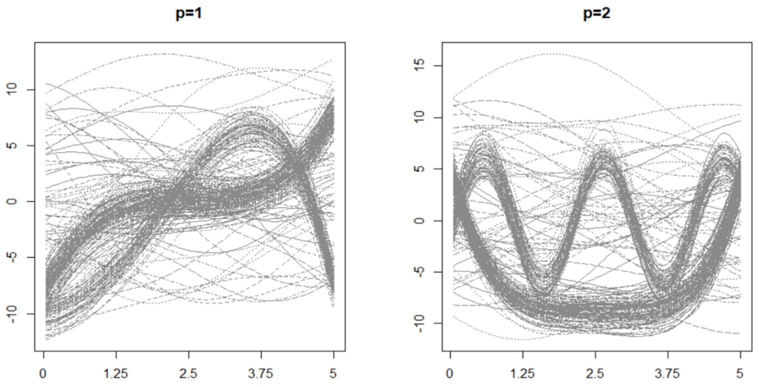

Figure 4: Obtained simulated data on the two considered dimensions

The HC2 algorithm is applied evaluating the scores using the norm defined in (4) with the L^2 norm on each p-dimensional Hilbert space and $w_p = 1$ for $p \in \{1,2\}$ and considering $(a,b) = (0,0)$. Parameters δ and θ are chosen as $\delta = 2$ and $\theta = 1$ after a sensitivity analysis, as explained in Section 3. The algorithm can easily reconstruct the bi-clustering structure: it finds three bi-clusters, represented by a different combination of the template functions on the two components, leaving all the other elements as not included in any bi-cluster.

Results are shown in Figure 5, where bi-cluster 0 represents the artificial bi-cluster containing the not assigned elements. Notice that, if we would have worked separately on each component, then the results would changed. Indeed, considering just the first dimension, the first and the last bi-clusters would have been merged together. Analogously, if the focus was on the second dimension, the second and the third bi-clusters would have belonged to the same bi-cluster. Thus, applying the HC2 bi-clustering we are able to unearth clear multivariate patterns which would be hidden if the analysis were conducted component-wise.

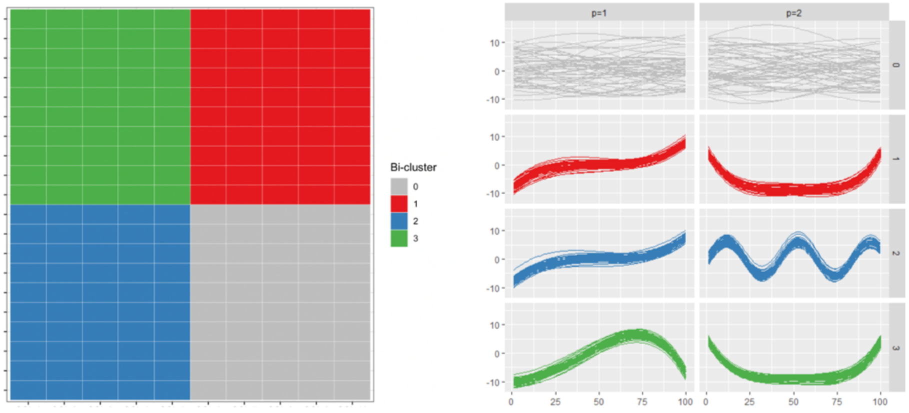

Figure 5: Obtained results applying the HC2 algorithm to the simulated data: resulting matrix (left), assigned functions to each bi-cluster (right)

5 Case-study: usage of a railway infrastructure

This section presents the results of the application of the methodological approach presented in previous sections to the collected data on the usage of the railway infrastructure in Lombardy on the Milano Cadorna - Asso line. Notice that, the developed approach has been thought to be applicable to study any line of the Lombardy railway infrastructure under examination, or more in general any railway system.

For each direction, we first build two data matrices focusing both on stations and trains as explained in Section 2, then, we perform the bi-clustering method presented in Section 3 on the obtained data matrices.

5.1 Bi-clustering stations over days

When observing the different stations along different days, we build a data matrix whose rows are the 19 stations and columns are the 9 considered days. Specifically, we construct two data matrices, each one for each direction of the trains on the line: from Milano Cadorna to Asso station and from Asso to Milano Cadorna station, respectively. For each data matrix, in each cell, representing a station in one day, we collect a three dimensional functional datum containing information about the trains load factor arriving and departing from that station, and the balance function of the boarding and dropping travellers in that station, along the day. The obtained functions are displayed in Figure 6; it can be noticed that the two directions are used by passengers differently according to the hour of the day, e.g. direction Milano Cadorna - Asso has a peak of load factor in the evening while Asso - Milano Cadorna has a peak in the morning. The HC2 algorithm, presented is Section 3, is then applied on this dataset with the aim of finding subgroups of stations behaving in a similar way over subgroups of days. To this purpose, we consider $(a, b) = (0, 0)$. Moreover, for the same reason, the L^2 norm is applied on each p-dimensional Hilbert space in the evaluation of the H-score and the rows/columns scores. The norm (4) is then employed setting $w_p = 1 \; \forall p \in \{1, 2, 3\}$, as all the considered functions have comparable units of measures. In this way, each bi-cluster is represented only by its average behaviour and the ideal bi-cluster is characterised by a group of functions all equal to each other, without considering any day or station effect within the same bi-cluster. A sensitivity analysis is performed before setting the value of θ and δ, as explained in Section 3, bringing to the choice of $\theta = 1$ and $\delta = 0.035$ for both directions.

The obtained results applying the HC2 algorithm on the two data matrices, corresponding to the different line directions, are reported in Figure 7 and 8. In Figure 7 the data matrices coloured by the obtained bi-clusters are reported; it can be noticed that the resulting bi-clusters are almost the same considering both directions, with a big bi-cluster covering stations from Asso to Meda during the whole week, and different bi-clusters on working days and weekends observing the stations from Seveso to Cesano and from Milano Affori to Milano Domodossola. Figure 8 shows for each bi-cluster its

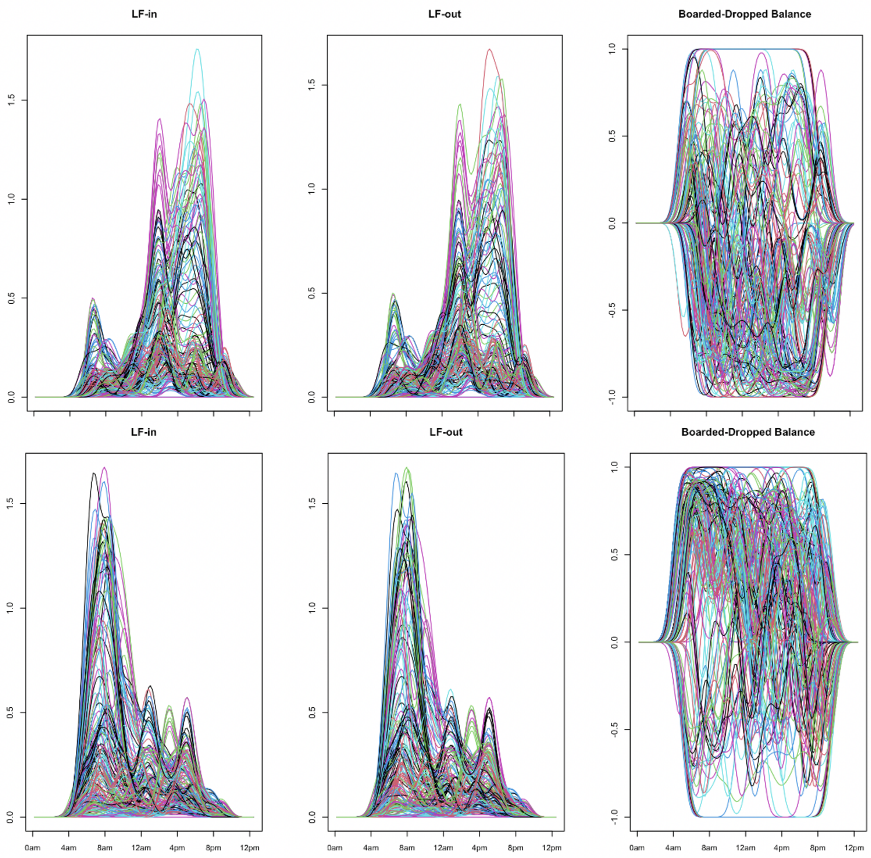

Figure 6: Obtained data for each station in each day for the directions Milano Cadorna - Asso (top) and Asso - Milano Cadorna (bottom)

template function.

In details, for the direction Milano Cadorna - Asso, the first bi-cluster, which covers about the 68% of the data, is related to all the stations from Meda to Asso during the whole week. It can be observed that trains arriving and departing from these stations along the whole day are uncrowded, and that dropping passengers are slightly more than boarding passengers during the whole day, thus, trains are slowly filling up in these stations. Bi-cluster 2 covers the stations from Milano Cadorna to Seveso during weekends, whose trains arriving and departing are uncrowded, and in which the number of boarding and dropping passengers is almost equal during the whole day. Bi-cluster 3 and bi-cluster 4 cover stations from Milano Domodossola to Milano

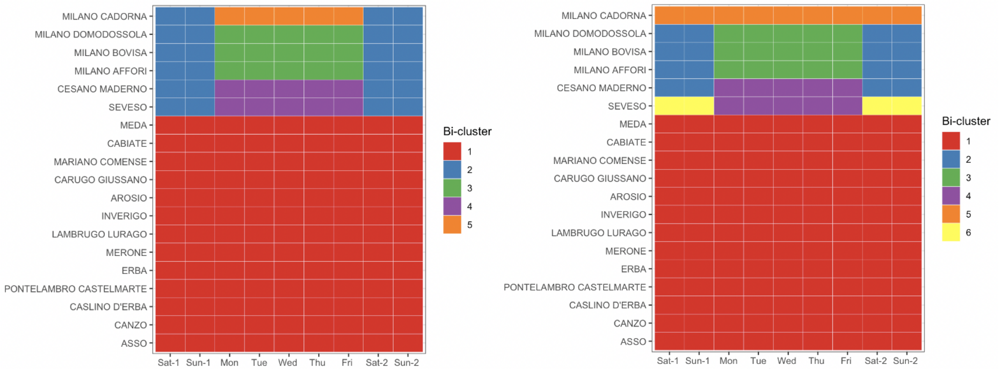

Figure 7: Obtained data matrices coloured by the different bi-clusters discovered applying the HC2 algorithm for the directions Milano Cadorna - Asso (left) and Asso - Milano Cadorna (right)

Affori and from Cesano to Seveso, respectively, during working days. Trains arriving and departing from these stations have a load factor higher than 100%, especially in the afternoon. The main difference among these two bi-clusters is that, especially in the afternoon, bi-cluster 3 covers stations in which passengers mainly take the train, i.e. the number of boarding passengers is higher than the number of dropping ones, while, as opposite, bi-cluster 4 covers stations in which passengers mainly drop off the trains, i.e. the number of dropping passengers is higher than the number of boarding ones. This aspect is evident also looking at the differences between the load factor in and out of bi-clusters 3 and 4: stations belonging to bi-clusters 3 have a LF of the arriving trains lower than the LF of the departing trains, meaning that passengers are boarding on the trains through these stations, hence increasing the trains crowding; as opposite, stations belonging to bi-clusters 4 have a LF of the arriving trains higher than the LF of the departing trains, meaning that passengers are dropping from the trains through these stations, hence decreasing the trains crowding.

Bi-cluster 5 covers Milano Cadorna station during working days: trains leaving this station have high crowding in the afternoon, around 100%.

Considering obtained bi-clusters in the opposite direction, the first bi-cluster, which covers about the 68% of the data, is related to all the stations from Asso to Meda during the whole week. It can be observed that trains arriving and departing in these stations along the whole day are uncrowded, and

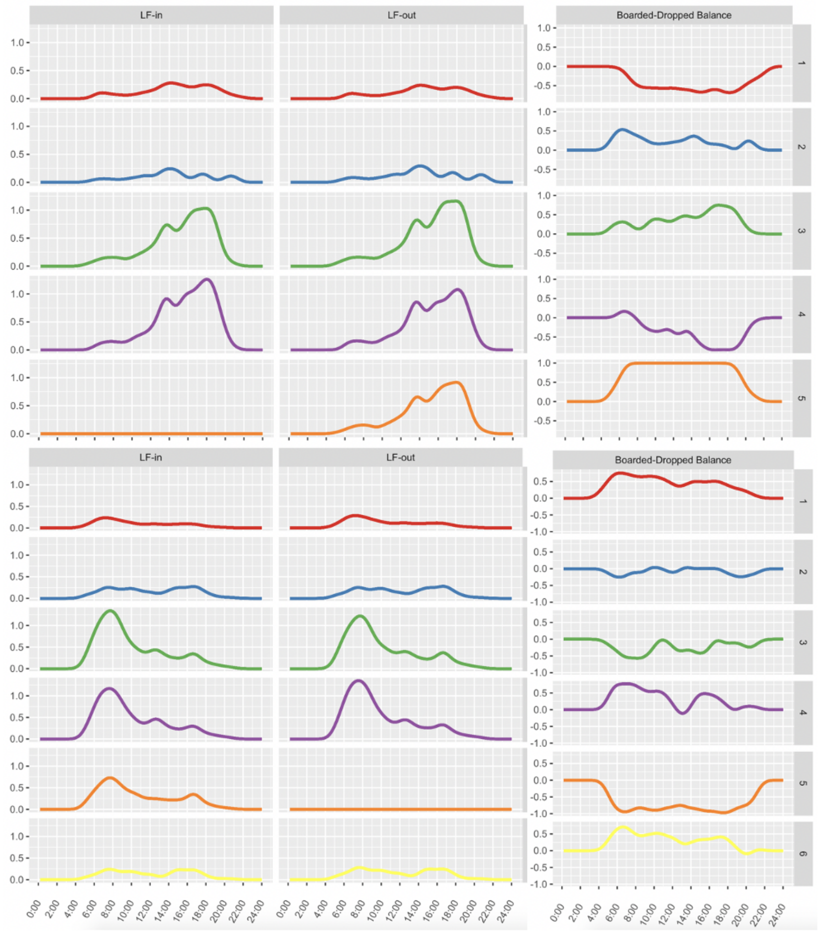

Figure 8: Obtained template functions for each bi-cluster on each dimension for the directions Milano Cadorna - Asso (top) and Asso - Milano Cadorna (bottom)

that there are more boarding than dropping passengers during the whole day. Bi-cluster 2 covers the stations from Seveso to Milano Domodossola during weekends, trains arriving and departing from these stations are uncrowded and the number of boarding and dropping passengers is almost equal during the whole day. Bi-cluster 3 and bi-cluster 4 cover stations from Milano Affori to Milano Domodossola and from Seveso to Cesano Maderno, respectively, during working days. In both the bi-clusters, trains arriving and departing from these stations have a load factor higher than 100%, especially in the morning. As in the obtained bi-clusters in the opposite direction, the main difference among these two bi-clusters is on the way passengers interact with the system through the covered stations; in this direction, passengers mainly drop off the trains through stations belonging to bi-cluster 3, while board on the trains through stations belonging to bi-cluster 4, especially in the morning. This aspect is also highlighted looking at the differences between the load factor in and out of the two bi-clusters: stations in bi-cluster 4 have a LF of arriving trains lower than LF of departing ones, the opposite for stations in bi-cluster 3. Bi-cluster 5 covers Milano Cadorna station during the whole week: trains arriving in this station have high crowding in the morning, around 80%, and here all the passengers drop off. Bi-cluster 6 covers Seveso station during the weekends. This station, as those in bi-cluster 2, has a low crowding, but differently from them has more boarding than dropping passengers during the whole day.

5.2 Bi-clustering of trains over days

As explained in previous sections, when observing the single trains travelling along the line in a specific day, we are building a data matrix whose rows are scheduled trains and columns are considered days. Specifically, we construct two data matrices whose dimensions change according to scheduled trains in a specific direction: from Milano Cadorna to Asso we constract a 22 x 9 data matrix, while from Asso to Milano Cadorna a 20 x 9. Notice that, trains are scheduled at different hours depending on the day of the week, so, when observing a data matrix, some cells contain missing values due to the fact that no trains are scheduled at that time in that day. In Figure 9 we show for each direction the data matrix structure highlighting the missing data.

For each cell of each data matrix, we collect three different information: the train load factor along the crossed stations, and the proportion of boarded and dropped passengers at each station with respect to the total travellers on the considered train. The obtained data for each train in each day in both directions are shown in Figure 10. The HC2 algorithm, presented is Section 3, is then applied on this dataset with the aim of finding subgroups of trains behaving in a similar way over subgroups of days. To this purpose, we consider $(a, b) = (0, 0)$. Moreover, for the same reason, the Euclidean norm is applied on each p-dimensional Hilbert space in the evaluation of the H-score and the rows/columns components. The norm (4) is then employed

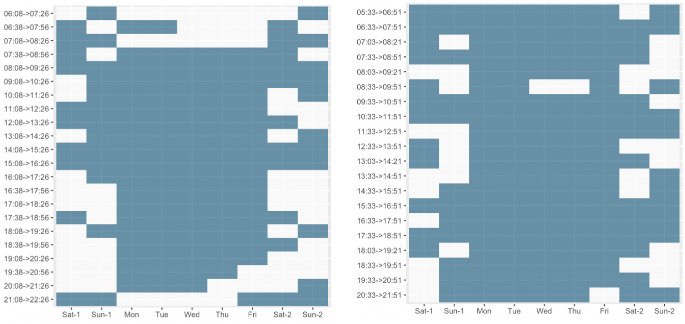

Figure 9: Scheduled trains during the different days (in blue) for the directions Milano Cadorna - Asso (left) and Asso - Milano Cadorna (right)

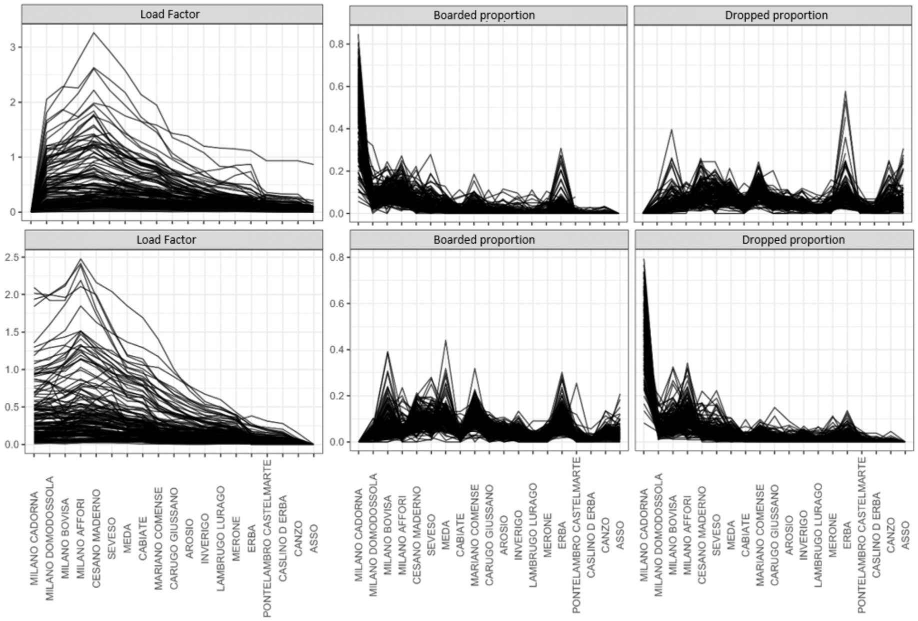

Figure 10: Obtained data for each train for the directions Milano Cadorna - Asso (top) and Asso - Milano Cadorna (bottom)

setting $w_p = 1$, for $p \in \{1, 2, 3\}$, as all the considered longitudinal vectors have comparable units of measures. In this way, each bi-cluster is represented only by its average behaviour and the ideal bi-cluster is characterised by a group of longitudinal data equal to each other, without considering any day or train effect within the same bi-cluster. A sensitivity analysis is performed before setting the value of θ and δ, suggesting a choice of $\theta = 1$ for both directions, $\delta = 0.009$ for the direction Milano Cadorna - Asso and $\delta = 0.008$ for the direction Asso - Milano Cadorna.

Obtained results on the two different directions are reported in Figure 11 and 12. In Figure 11 we show the obtained data matrices coloured by found bi-clusters. Note that different bi-clusters are observed when considering the two different directions. In detail, when considering the trains travelling from Milano Cadorna to Asso four different bi-clusters are discovered, whose behaviour, with respect to each dimension, is shown in Figure 12 (top). Bi-cluster 1 covers trains before 1:08p.m., between 3.08p.m. and 4:38p.m. and after 7:38p.m. during the whole week, which are uncrowded along the whole line with more dropped in Erba station. Bi-cluster 2 covers trains at 2:08p.m., 5:08p.m., 6:08p.m. and 7:08p.m. during the whole week except Thursday, which are overcrowded (with load factor higher than 100%), especially in the first part of the line, and there are more dropped in Cesano Maderno and Seveso stations. Bi-clusters 3 and 4 cover the same trains as bi-cluster 2 but on Thursday, which are overcrowded (with load factor respectively higher than 150% and 200%).

When observing the opposite direction, the structure of found bi-clusters is different. Firstly, bi-clusters containing crowded trains are almost in the morning and the bi-clustering structure highlights a difference between working days and weekends. In details, four bi-clusters are discovered: bi-cluster 1, covering trains before 6a.m. and after 9:30a.m. during the whole week, which are uncrowded along the whole line and where there are more boarded passengers in Milano Bovisa station with respect to other bi-clusters; bi-cluster 2, covering trains between 6:30a.m. and 8:30a.m. during the weekends, which are uncrowded along the whole line; bi-clusters 3 and 4, covering the same scheduled trains as in Bi-cluster 2 but during working days, which contain overcrowded trains (with load factors respectively higher than 150% and 200%) especially in the last part of the line.

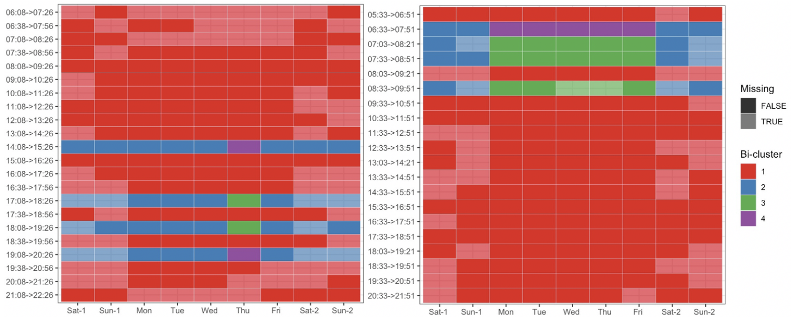

Figure 11: Obtained data matrices coloured by the different bi-clusters discovered applying the HC2 algorithm for the direction Milano Cadorna - Asso (left) and Asso - Milano Cadorna (right)

5.3 General Results on the Cadorna-Asso line

Results presented above have highlighted situations of overcrowding and have shown how passengers travel on the Cadorna-Asso line through two different approaches: considering stations or trains over different days. Even if the two approaches are different, they point out similar results and display complementary information for an insightful understanding of the dynamics of the infrastructure. Generally, the line seems to be used mainly by commuters during working days in both directions: in the morning trains going to Milano Cadorna present high events of crowding in the stations just before the gates of Milan and in the city itself, while, on the opposite direction, trains present high events of crowding in the same stations but in the afternoon.

The obtained bi-clusters from a station point of view suggest that a different behaviour is present between stations in the city center, Cesano Maderno and Seveso during working days and weekends, and stations from Meda to Asso along the whole week. In particular a problem of crowding is identified in stations closed to Milan. These results are underlined in both directions, despite the fact that different behaviours are present according to the observed direction.

The obtained bi-clusters from a train point of view suggest the scheduled trains which need some improvement in capacity due to overcrowding problems and trains that can be deleted, as uncrowded. Overcrowding issues are identified at different hours of the day according to the direction: in the afternoon from Milano Cadorna to Asso and in the morning in the op-

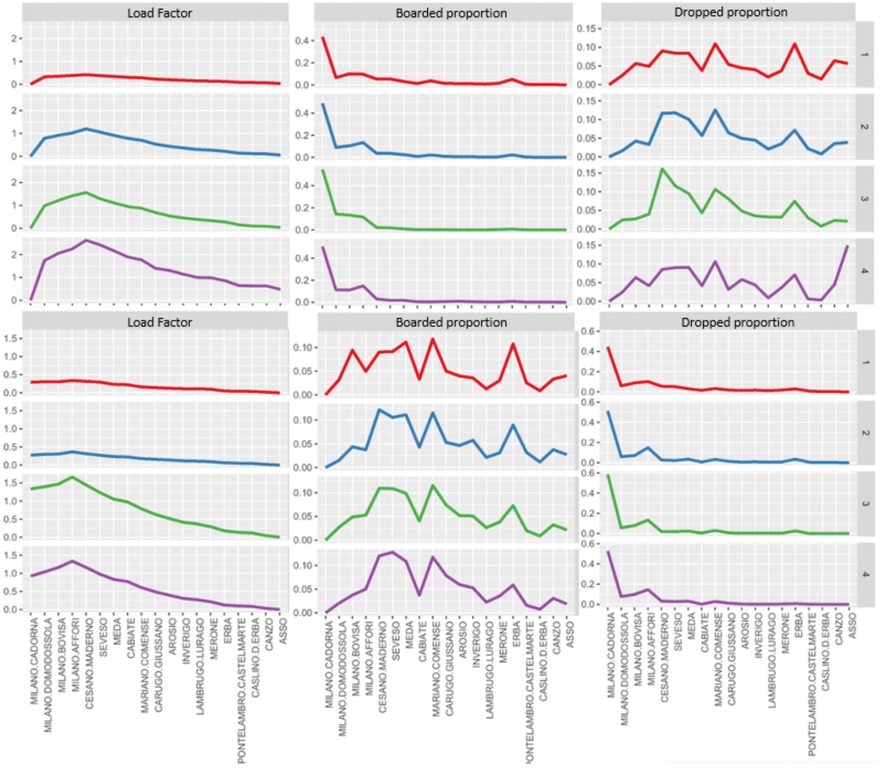

Figure 12: Obtained template functions for each bi-cluster on each dimension for the direction Milano Cadorna - Asso (top) and Asso - Milano Cadorna (bottom)

posite direction. Observed proportion of boarding and dropping on specific bi-clusters could guide possible decisions on improvement or reduction of the number of trains in the line, considering improvement only for some stops - e.g. skipping a station or increasing the frequency of trains in a section of the line.

6 Conclusions

In this work we study the passengers flow on the lines of the Lombardy railway infrastructure analysing data collected through people counter technologies. The aim of this work is to develop a strategy to identify situations of crowding in the system and to underlay the spatio-temporal interactions

of passengers with the infrastructure, so to identify the potential weaknesses of the public transport system and help decision makers in improving the service quality.
For each station the load factor of arriving and departing trains and a balance function of boarded and dropped are considered as functional data along time. Similarly, for each train, the load factor and the proportion of boarded and dropped at each station are evaluated as longitudinal vectors along the stops of a line. These data are observed over a period of nine days with the aim of discovering subgroup of stations (or trains) behaving in a similar way over subgroups of days, observing the different collected information. To this end, a bi-clustering technique - called Hilbert Chung and Church (HC2) - is presented to group simultaneously rows and columns of a data matrix, whose elements are complex object data embedded in a Hilbert space. The presented approach is non parametric and very flexible, allowing to discover different bi-clusters depending on the problem at hand.
Applying the HC2 algorithm to the data at hand we are able to discover how passengers move along a line underlying situations of crowding in the system depending on the day of the week and the hour of the day. The two different points of view employed in the analyses, i.e. bi-clustering respectively stations and trains along days, allow to obtain a complete understanding of the system. The obtained results could help railway transport companies to study the system, identify eventual issues and plan eventual decisions on the trains scheduling. The developed approach is flexible and scalable, indeed it is ready to be used to analyse larger datasets and different railway systems in other region.

Acknowledgement

The authors gratefully acknowledge Trenord for providing the data and for part of the financial support of Dr. Marta Galvani. We also thank our colleagues prof. Giovanni Azzone, prof. Marika Arena and dr. Valeria Urbano for stimulating and fruitful discussions.

References

[1] Understanding attitudes towards public transport and private car: A qualitative study, Transport Policy 14 (6) (2007) 478 – 489.

[2] I. Ceapa, C. Smith, L. Capra, Avoiding the crowds: understanding tube station congestion patterns from trip data.

[3] J.-M. Chiou, et al., Dynamical functional prediction and classification, with application to traffic flow prediction, The Annals of Applied Statistics 6 (4) (2012) 1588–1614.

[4] I. G. Guardiola, T. Leon, F. Mallor, A functional approach to monitor and recognize patterns of daily traffic profiles, Transportation Research Part B: Methodological 65 (2014) 119–136.

[5] F. Crawford, D. Watling, R. Connors, A statistical method for estimating predictable differences between daily traffic flow profiles, Transportation Research Part B: Methodological 95 (2017) 196–213.

[6] A. Torti, A. Pini, S. Vantini, Modelling time-varying mobility flows using function-on-function regression: Analysis of a bike sharing system in the city of milan, Journal of the Royal Statistical Society: Series C (Applied Statistics) 70 (1) (2021) 226–247.

[7] J. S. Marron, A. M. Alonso, Overview of object oriented data analysis, Biometrical Journal 56 (5) (2014) 732–753.

[8] C. Bouveyron, E. Côme, J. Jacques, et al., The discriminative functional mixture model for a comparative analysis of bike sharing systems, The Annals of Applied Statistics 9 (4) (2015) 1726–1760.

[9] E. Thuillier, L. Moalic, S. Lamrous, A. Caminada, Clustering weekly patterns of human mobility through mobile phone data, IEEE Transactions on Mobile Computing 17 (4) (2017) 817–830.

[10] A. Godichon-Baggioni, C. Maugis-Rabusseau, A. Rau, Clustering transformed compositional data using k-means, with applications in gene expression and bicycle sharing system data, Journal of Applied Statistics 46 (1) (2019) 47–65.

[11] L. He, B. Agard, M. Trépanier, A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method, Transportmetrica A: Transport Science 16 (1) (2020) 56–75.

[12] Y. Cheng, G. M. Church, Biclustering of expression data., in: Ismb, Vol. 8, 2000, pp. 93–103.

[13] Y. B. Slimen, S. Allio, J. Jacques, Model-based co-clustering for functional data, Neurocomputing 291 (2018) 97–108.

[14] C. Bouveyron, L. Bozzi, J. Jacques, F.-X. Jollois, The functional latent block model for the co-clustering of electricity consumption curves, Journal of the Royal Statistical Society: Series C (Applied Statistics) 67 (4) (2018) 897–915.

[15] G. Govaert, M. Nadif, Co-clustering: models, algorithms and applications, John Wiley & Sons, 2013.

[16] M. Galvani, A. Torti, A. Menafoglio, S. Vantini, Funcc: a new biclustering algorithm for functional data with misalignment, Computational Statistics Data Analysis.

[17] J. Di Iorio, S. Vantini, funbi: a biclustering algorithm for functional data, MOX-Report No. 46/2019.

[18] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

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

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

20/2021 Pasquale, A.; Ammar, A.; Falcó, A.; Perotto, S.; Cueto, E.; Duval, J.-L.; Chinesta, F.
A separated representation involving multiple time scales within the Proper Generalized Decomposition framework

16/2021 Salvador, M.; Dede', L.; Manzoni, A.
Non intrusive reduced order modeling of parametrized PDEs by kernel POD and neural networks

17/2021 Chew, R.; Benacchio, T.; Hastermann, G.; Klein, R.
Balanced data assimilation with a blended numerical model

18/2021 Gigante, G.; Vergara, C.
On the choice of interface parameters in Robin-Robin loosely coupled schemes for fluid-structure interaction

19/2021 Gillard, M.; Benacchio, T.
FT-GCR: a fault-tolerant generalized conjugate residual elliptic solver

13/2021 Ferro, N.; Perotto, S.; Cangiani, A.
An anisotropic recovery-based error estimator for adaptive discontinuous Galerkin methods

14/2021 Peli, R.; Menafoglio, A.; Cervino, M.; Dovera, L.; Secchi, P;
Physics-based Residual Kriging for dynamically evolving functional random fields

15/2021 Fumagalli, A.; Patacchini, F.S.
Model adaptation in a discrete fracture network: existence of solutions and numerical strategies

12/2021 di Cristofaro, D.; Galimberti, C.; Bianchi, D.; Ferrante, R.; Ferro, N.; Mannisi, M.; Perotto, S
Adaptive topology optimization for innovative 3D printed metamaterials

11/2021 Antonietti,P.F.; Manzini, G.; Mazzieri, I.; Scacchi, S.; Verani, M.
The conforming virtual element method for polyharmonic and elastodynamics problems: a review