A novel spatio-temporal clustering technique to study the bike sharing system in Lyon

Marta Galvani*
University of Pavia
Pavia, Italy
marta.galvani@unipv.it

Agostino Torti†
Politecnico di Milano
Milan, Italy
agostino.torti@polimi.it

Alessandra Menafoglio‡
Politecnico di Milano
Milan, Italy
alessandra.menafoglio@polimi.it

Simone Vantini§
Politecnico di Milano
Milan, Italy
simone.vantini@polimi.it

ABSTRACT

In the last decades cities have been changing at an incredible rate growing the needs of efficient urban transportation to avoid traffic jams and high environmental pollution. In this context, bike sharing systems (BSSs) have been widely adopted by major regions like Europe, North America and Asia-Pacific becoming a common feature of all metropolitan areas. Its fast growing has increased the need of new monitoring and forecasting tools to take fast decisions and provide an efficient mobility management.

In this context we focus on the BSS of Lyon in France, called Vélo'v. In particular we analyse a dataset containing the loading profiles of 345 bike stations over one week, treating the data as continuous functional observations over a period of one day. The aim of this work is to identify spatio-temporal patterns on the usage of bike sharing stations, identifying groups of stations and days with similar behaviour, with the purpose of providing useful information to the fleet managers. To this scope, we develop a novel bi-clustering algorithm able to deal with functional data, extending a nonparametric algorithm developed for multivariate data. This new algorithm is able to find simultaneously subsets of rows and columns with similar behaviour when the elements of the dataset are functional objects.

Obtained results show that through this analysis it is possible to identify different usage patterns according to the area of the city and the day of the week.

1 INTRODUCTION

Due to urbanization and globalization in the last decades, cities have been changing at an incredible rate. In particular a growing need of urban transportation has increased the number of vehicle usage on roads and ultimately led to heavy traffic jams and high environmental pollution. To alleviate the above growing issues, the bike sharing program has been widely adopted by major regions like Europe, North America and Asia-Pacific.

Bike sharing systems (BSSs) have become a common feature of all metropolitan areas and according to a 2019 Global Market Insights, Inc. report, it has been predicted that the fleet size of bike sharing market will gain over 8% from 2019 to 2025, leading the worldwide industry revenue to surpass a valuation of USD 10 billion.

This fast growth has urged scientists in developing suitable monitoring and forecasting tools to handle with mobility management and make feasible and efficient future plans [5]. Many studies have demonstrated that an efficient analysis of the data collected by BSSs can provide good insights for the service design, i.e. for the reallocation strategies optimization, to underly the causes of network imbalance and for the adjustment of pricing policies ([10], [14], [1]).

Most BSSs provide open access to their data regarding the real-time status on their bike stations. In this context we focus on the BSS of Lyon, called Vélo'v. Launched in 2005, Vélo'v is the first bicycle-sharing system in France, with a network of more than 3000 bikes spread over 345 stations in Lyon and neighboring Villeurbanne. The service has been developed by street furniture company JCDecaux for Lyon Metropole and it counts now more than 68.500 subscribers.

In this work we analyse a dataset containing the loading profiles of 345 bike stations over one week during the period from Monday 10th March until Sunday 16th March in 2014. The real time data are available at https://developer.jcdecaux.com/ trough an api key and they were first used in [3]. Specifically, for each station the number of available bikes divided by the total number of bike docks at each hour is recorded.

The aim of our work is to understand the spatio-temporal patterns of the bike stations usage, providing useful information for the correct management of the service. We are interested in understanding how bike sharing stations are used according to their spatial position looking at the variability within and between days.

Due to the continuous dependence on time of our data, we decide to model them making use of tools from Functional Data Analysis (FDA), the branch of statistics dealing with curves, surfaces or anything else varying over a continuum (e.g., [13]). In this way it is possible to consider the within-day variability.

From a statistical point of view, we are facing with a problem of clustering both the bike stations and the days of the week, which is know in the literature as a bi-clustering problem. The main aim

[*] Department of Mathematics
[†] MOX Laboratory for Modeling and Scientific Computing - Department of Mathematics.
Center for Analysis Decisions and Society, Human Technopole.
[‡] MOX Laboratory for Modeling and Scientific Computing - Department of Mathematics.
[§] MOX Laboratory for Modeling and Scientific Computing - Department of Mathematics.

of bi-clustering (or co-clustering) algorithms is to simultaneously group individuals and features into homogeneous sets, see [11] for a complete review of bi-clustering methodologies.

As for each station and for each day we define the bike station loading profile as a continuous functional datum, we have found ourselves with a problem of bi-clustering functional data. Different methodologies, which extend some well known algorithm for clustering multivariate data, have been proposed in the literature with the aim of clustering functional data [9]. In the same way, bi-clustering methods can be generalized to functional data by defining new algorithms able to detect functional bi-clusters. Although the concept of bi-clustering was first introduced by [7] in the 1970s, [4] are recognized as the first ones to propose a bi-clustering algorithm. Subsequently different model based approaches have been proposed, among them [6] relies on the latent block model. Starting from it, [2] introduce a parametric model able to deal with functional data. Although, as based on Latent block model, this approach is able to detect exhaustive bi-clustering maintaining a checkerboard structure that does not always fit with the real situations.

In this work we introduce a novel methodology based on the extension of the Cheng and Church algorithm with the aim of detecting functional non exclusive bi-clusters. We propose an iterative procedure based on a non parametric approach obtaining a deterministic strategy that does not have to rely on strong modelling assumptions of the data, which are generally not consistent in the FDA framework, and allows for flexible non exclusive bi-clusters.

The rest of this paper is organized as follows: in Section 2 we describe the functional Cheng and Church bi-clustering, discussing how the extension of the algorithm is implemented. In Section 3 the introduced methodology is applied on the Vélo'v BSS and the main results are reported. In Section 4 conclusions are presented and discussed, underlying the spatio-temporal patterns found in the data employing the novel algorithm proposed in this work.

2 METHODOLOGY: THE FUNCTIONAL CHENG AND CHURCH ALGORITHM

Given a dataset arranged in a matrix A composed by n rows and m columns, the aim of a bi-clustering technique is to find a submatrix $A' \in A$, corresponding to a subset of rows I and a subset of columns J, with a high similarity score. In particular, in the Cheng and Church algorithm ([4]), an *ideal* bi-cluster is a set of rows I and a set of columns J such that each element a_{ij} in the bi-cluster can be expressed as $a_{ij} = a_{IJ} + \alpha_i + \beta_j \; \forall i \in I$ and $\forall j \in J$, where a_{IJ} is the average value in the bi-cluster and α_i and β_j are respectively the residue of rows and columns average value and the total average value a_{IJ}. A particular measure of goodness is evaluated for a bi-cluster $A'(I,J)$ considering a score H which is the *Mean Squared Residue* between all the real values $a_{ij} \in A'(I,J)$ and their representative values in the bi-cluster $a_{IJ} + \alpha_i + \beta_j$.

Extending these concepts to FDA, each element of the dataset matrix A is a function $f_{ij}(t)$ defined on a continues domain T. In such framework we define an ideal bi-cluster A' as a subset of rows I and columns J such that each function belonging to the bi-cluster $A'(I,J)$ can be defined as follows:

$$f_{ij}(t) = f_{IJ}(t) \; \forall i \in I \text{ and } \forall j \in J$$

where $f_{IJ}(t) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} f_{ij}(t)$. For easiness of interpretation we define the bi-cluster template observing only the average function in the bi-cluster.

Consequently, the extended H-score of a bi-cluster $A'(I,J)$ is:

$$H_{IJ} = \int_T H_{IJ}(t)$$

with $H_{IJ}(t) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} \left\{ f_{ij}(t) - [f_{IJ}(t)] \right\}^2$.

The pseudo-code of the Algorithm to find a bi-cluster works as expressed in algorithm 1. The algorithm starts considering the whole dataset and try to find the biggest bi-cluster with a H-score value lower then a given threshold δ. The rows/columns addition and deletion procedures are a natural extension of the ones introduced in [4]. The procedure follows the main structure of the original Cheng and Church algorithm, except for the masking procedure. Indeed, instead of this step, after finding a new bi-cluster, a set of all the biggest submatrices containing elements not already assigned is found through the Bimax algorithm ([12]). Then, in the next iteration, the new bi-cluster is searched inside the biggest submatrix found. Each time a new bi-cluster is found the set of the submatrices of not assigned elements is updated; otherwise a new bi-cluster is searched in the following biggest submatrix in the set.

Algorithm 1: Functional Cheng and Church algorithm

```
Input: $(n, m)$ matrix $A$ whose elements are functions
       $f_{ij}(t)$
       $\delta$ >0 the maximum acceptable $H$-score
       $maxIter$ the maximum number of allowed
       iterations
Result: A set of Bi-clusters with $H$-score< $\delta$
Set M=A
while $iter < maxIter$ and submatrices to search in for
  bi-clusters are present do
    Given a submatrix M:
    while $H$-score> $\delta$ and deletion/addition is still possible
     do
        (1) Multiple Node Deletion:
            remove groups of rows/cols
        (2) Single Node Deletion:
            remove the row/col that reduce $H$-score
            the most
        (3) Node Addition:
            add rows/cols that do not make $H$-score
            greater than $\delta$
    end
    if A new bi-cluster is found then
        Apply Bimax to search for all the biggest
         submatrices of not assigned elements and select
         the biggest one as M
    else
        Select as M the following biggest submatrix
    end
end
```

As in the classical Cheng and Church, the results are sensitive to the choice of the input parameter δ. Indeed, a too high value of δ would imply a unique big bi-cluster, while a too low value would imply a really large number of bi-clusters or even the

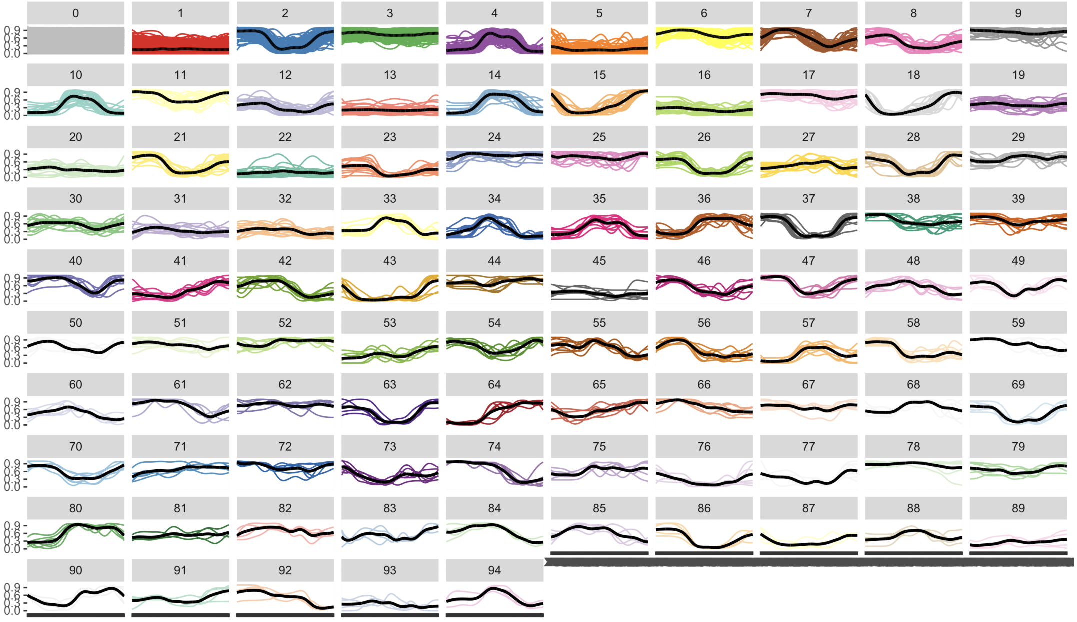

Figure 1: Functions belonging on each bi-cluster with their templates

impossibility to find a bi-cluster. To tune the parameter δ, we perform a sensitivity analysis on the number of obtained clusters, the number of not assigned observations and the sum of the H scores over all the found bi-clusters. Then, following the same approach used for many other clustering techniques as the classical k-means, we choose the value of δ where an elbow or a peak are found.

3 DATA ANALYSIS AND RESULTS

The first step of our analyses is to treat the available raw data as continuous functions. Specifically, for each station and for each day we define, through a kernel density estimation smoothing method [8], the bike station loading profile as a continuous functional datum representing the number of available bikes divided by the total number of bike docks at each timestamp. In this way we obtain 2415 curves, i.e. 345 stations per 7 days, representing all the elements $f_{ij}(t)$ (with $t \in [0, 24]$) of a dataset matrix A with the same dimensions (345x7). Functional Cheng and Church algorithm, presented in the previous section, can be applied on this dataset.

To find the set of the best bi-clusters, a threshold δ must be provided to the algorithm. After performing a sensitivity analysis to choose the threshold parameter δ we fix δ equal to 0.025.

Results are shown in Figure 1. There are in total 94 bi-clusters while the not assigned observations are artificially assigned to bi-cluster 0. For each bi-cluster all the functions contained in that bi-cluster are shown together in colors; the template function, defined as the average curve of the bi-cluster, is displayed in black. Looking at the bi-cluster dimensions (i.e. the number of curves in the bi-cluster), the obtained results are able to explain the 75% of the data, while the 25% of the functions are not assigned to any bi-clusters. Note that the found bi-clusters have been ordered from the biggest one to the smallest one, considering the number of included rows and columns. Evaluating the percentage of working and weekend days for each bi-cluster, we notice that some bi-clusters cover specific patterns of the working days or of the weekends (e.g. bi-clusters 4, 5, 6), while some other considered stations that have the same pattern during working and weekend days (e.g. bi-clusters 1, 15, 18).

Turning our attention on the found bi-clusters, Figure 1, it is possible to interpret results as a way of segment the city in different activity areas according to the day of the week and the hour of the day. Observing the usage profiles of the bi-clusters, three main groups can be identified: the *constant profile*, the *residential profile* and the *working profile*.

The *constant profile* bi-clusters show flat functions of usage during the whole day implying a no usage or a continuous replacement of bikes in these stations. Among these, the always Full (e.g. bi-clusters 3, 6 and 9) and Empty stations (e.g. bi-clusters 13, 16 and 20), which necessarily imply user dissatisfaction as they respectively cannot drop-off or pick-up a bike, are of particular interest.

The *working profile* (e.g. bi-clusters 4, 10 and 14) and the *residential profile* (e.g. bi-clusters 2 and 37) instead, are characterised by a huge activity during rush hours in the morning, around 8a.m., and in the evening, around 7p.m.. However, the two groups show an opposite behaviour since while the first one fills up in the morning and empties out in the evening, instead, the second one empties out in the morning and fills up in the evening. Moreover, looking at the days inside the *working profile* and *residential profile* groups, it appears that these bi-clusters are composed by working days. The peculiarity of these two groups reveal a clear commuting behaviour of the bike sharing users which move during working days in the morning and evening rush hours.

To better explore these two behaviours, we focus, as explanatory example, on bi-clusters 2 and 4. In Figure 2 and 3 results on these two bi-clusters are reported; in particular all the functions belonging to the bi-cluster with the bi-cluster template (in black), the corresponding days and bike stations location are shown. Observing Figure 2 it is possible to say that Bi-cluster 2 is a block composed from 34 stations and 5 days (from Monday to Friday), covering almost the 7% of all the data. It is characterised from full stations before 8a.m. and after 8p.m. and empty stations during

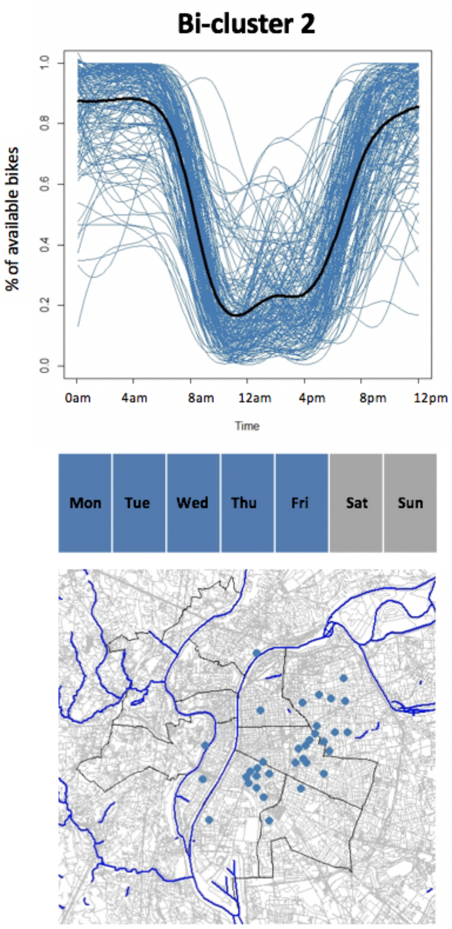

Figure 2: Functions in bi-cluster 2 with calendar and bike stations position on the map of Lyon

the rest of the day. This peculiar behaviour is justified by the fact that these stations are mostly located in residential areas in the East of the city. An opposite behaviour is instead present in all the stations belonging to bi-cluster 4 (Figure 3) which are full between 8a.m.-8p.m. and empty in the rest of the day. This behaviour is easily explainable by the fact that these stations are located in parts of the city with many companies where people are used to commute during the day. This bi-cluster is composed by 17 stations and again the 5 working days from Monday to Friday, covering the 3.5% of the total observations in the data.
Another small group of bi-clusters, almost covering weekend days, can be described as *weekend profile* (e.g. bi-clusters 6, 7 and 73). For instance, bi-cluster 73 (Figure 4) contains the daily usage profiles of 3 stations for the entire weekend. The peculiarity of this bi-cluster is that the concerned bike stations are in the city center, very closed to River Sâone banks, where there are many shops and bars especially active during the weekends. It is possible to see, observing Figure 4, that these stations are filled up during evening until they become almost totally full before midnight and then they slowly empty out during the night. This behaviour can be explained considering that people go out clubbing during evening and then they return back home late in the night.

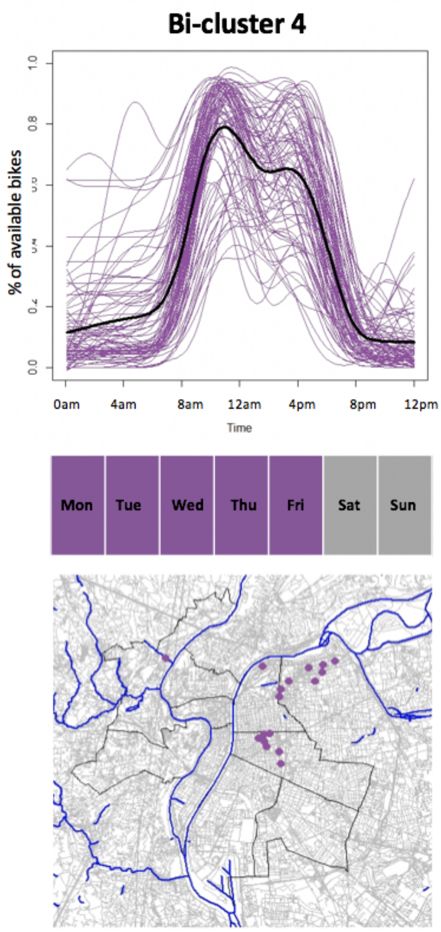

Figure 3: Functions in bi-cluster 4 with calendar and bike stations position on the map of Lyon

4 CONCLUSION

The aim of our work is to study the spatio-temporal patterns of the Vélo'v BSS usage profile during a one week period in Lyon providing useful information to the fleet managers. To this end, we model the usage profiles of the different bike stations around the city day by day as continuous functions with the aim of discovering subgroups of stations and days with similar behaviour, which is know in the literature as a bi-clustering problem.
To build our analyses, we introduce a novel non parametric bi-clustering algorithm extending the Cheng and Church algorithm in the FDA framework. From a methodological point of view the concept of ideal bi-cluster is extended for functional data and a new score evaluation for found bi-clusters is introduced. In addition the new introduced algorithm overcomes the main weaknesses of the original Cheng and Church, avoiding the usage of the masking procedure and introducing a greedy search in the not already assigned elements. The developed algorithm allows to find non exclusive bi-clusters with a H-score lower than a given value δ, through a non parametric procedure. This has the advantage of avoiding to rely on strong modelling. A sensitivity analysis for the δ parameter tuning is also presented.

From a practical point of view, the developed approach is applied to study the daily usage profiles of all the 345 stations of the Vélo'v BSS in Lyon for one week in March 2014. Results show

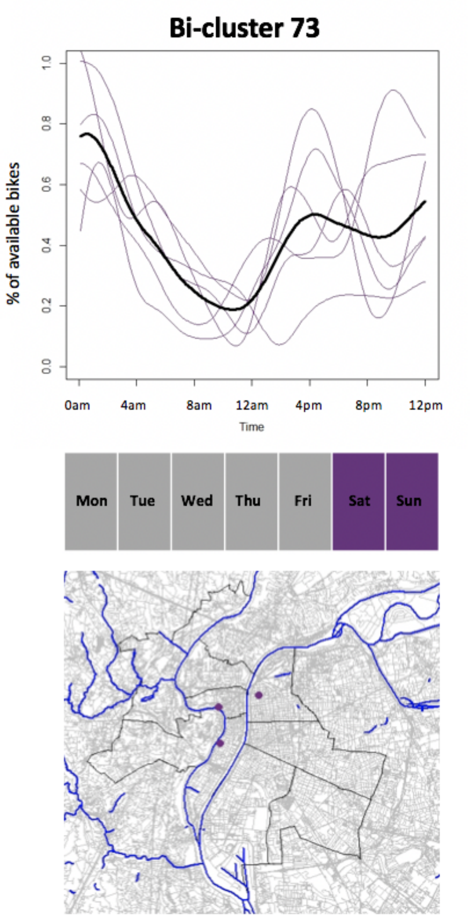

Figure 4: Functions in bi-cluster 73 with calendar and bike stations position on the map of Lyon

clear patterns of usage allowing to segment the city in different activity areas according to the day of the week and the hour of the day. For instance, a commuting behaviour is observed revealing that stations next to residential areas and working areas have an opposite behaviour during working days. It is interesting to notice that despite no apriori information about the spatial distribution of the stations are taken into account by the model, it appears that stations belonging to the same bi-cluster are actually located in neighborhoods with the same socio-economic characteristics. Moreover, groups of stations always full or always empty are highlighted, revealing some criticalities of the service.

In conclusion, our work contributed to implement the study of a bike sharing system in two ways: from a methodological point of view, we defined a novel non parametric bi-clustering technique for functional data; from an applied point of view, we analysed the bike sharing system in the city of Lyon providing useful information for the correct management of the service.

REFERENCES

[1] Pierre Borgnat, Celine Robardet, Jean-Baptiste Rouquier, Patrice Abry, Patrick Flandrin, and Eric Fleury. 2011. Shared Bicycles in a City: A Signal Processing and Data Analysis Perspective. *Advances in Complex Systems* 14 (06 2011). https://doi.org/10.1142/S0219525911002950

[2] Charles Bouveyron, Laurent Bozzi, Julien Jacques, and François-Xavier Jollois. 2018. 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.

[3] Charles Bouveyron, Etienne Côme, Julien Jacques, et al. 2015. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics* 9, 4 (2015), 1726–1760.

[4] Yizong Cheng and George M Church. 2000. Biclustering of expression data.. In *Ismb*, Vol. 8. 93–103.

[5] Elliot Fishman. 2016. Bikeshare: A Review of Recent Literature. *Transport Reviews* 36, 1 (2016), 92–113. https://doi.org/10.1080/01441647.2015.1033036

[6] Gérard Govaert and Mohamed Nadif. 2013. *Co-clustering: models, algorithms and applications.* John Wiley & Sons.

[7] John A Hartigan. 1972. Direct clustering of a data matrix. *Journal of the american statistical association* 67, 337 (1972), 123–129.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning.* Springer New York Inc., New York, NY, USA.

[9] Julien Jacques and Cristian Preda. 2014. Functional data clustering: a survey. *Advances in Data Analysis and Classification* 8, 3 (2014), 231–255.

[10] Neal Lathia, Saniul Ahmed, and Licia Capra. 2012. Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies* 22 (2012), 88 – 102. https://doi.org/10.1016/j.trc.2011.12.004

[11] Beatriz Pontes, Raúl Giráldez, and Jesús S Aguilar-Ruiz. 2015. Biclustering on expression data: A review. *Journal of biomedical informatics* 57 (2015), 163–180.

[12] Amela Prelić, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22, 9 (2006), 1122–1129.

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

[14] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. 2011. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences* 20 (2011), 514 – 523.