Smart Statistics for Smart Applications

Book of Short Papers SIS2019

SIS 2019
SMART STATISTICS FOR SMART APPLICATIONS

SIS Società Italiana di Statistica

Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

Published by Pearson

www.pearson.com

Giugno 2019 ISBN 9788891915108

Preface

Section 1. Plenary Sessions and Round Table

Section 2. Invited Papers

Section 3. Contributed Papers

Section 4. Posters

Preface

Preface

This book includes the scientific contributions presented at the Intermediate Meeting of Italian Statistical Society (SIS) held in Milan at the Universit Cattolica del Sacro Cuore, from June 18th to 21th of 2019. Following a long tradition (and a statutory indication of the Society), the intermediate meetings are held bi-annually on specific themes. This year, aiming at bridging the gap between statistics and the world of Big Data and Data Science, the conference was entirely devoted to the theme of "Smart Statistics for Smart Applications". In this way the Italian Statistical Society had the explicit intention to answer the high and rapidly increasing demand on the subject, by providing academics, researchers and practitioners with a forum where new ideas and new methods could meet with new needs, new research questions and new applications.

The Conference could not have been organized without the joint effort of the Milanese network of Universit Cattolica del Sacro Cuore, Universit degli Studi di Milano Bicocca, Universit Bocconi, Universit "Vita e salute" San Raffaele, Politecnico di Milano and Universit Statale di Milano. Members of all these universities took part actively to the Local Organizing Committee. The Conference has also greatly benefited from the contribution of the strategic partner Mathesia, which contributed to the various aspects of the organization, with special focus to the active involvement of private firms and companies and of the non-academic components.

The conference has registered more than 200 scientific contributions, including papers presented in plenary invited sessions, papers collected in specialized and solicited sessions on specific themes, about 100 contributions spontaneously submitted to the Program Committee and a poster session. All contributions were focused on the conference theme and provided a good overview of the state-of-the-art of the subject, from methodological and theoretical contributions, to applied works and case studies. The two plenary lectures were devoted to the (provocative) idea of "shallow learning", as opposed to the more in-vogue idea of deep learning", and to the problems linked with Big Data veridicity and reliability. A plenary round table draw the participants attention on the concept of smart ageing.

A distinctive feature of this conference, relative to previous analogous experiences, was the presence of many round tables and activities focused on topics of interest for a wider audience, freely open to external participation. These activities were termed Fuoriconvegno" and included a special session on "Data skills: Statistics and education for future jobs" organized jointly with Pearson Italia Publisher, a round table on "How to Close the Gap Between the Practice and Theory in Digital Transformation Era" organized joint with Mathesia, a colloquium on "Big Data and Big Responsibility", a round table on "Political polls in the Big Data era", a round table on "Big Data and Public Administration", a round table on the changing role of the statistical scientific societies in a new interconnected world and the fifth edition of the statistical competition "Stats Under the Stars (SUS5)" organized by the Bocconi University, a whole-night hackathon on real-world business analytic problems for young Data Scientists.

More information about the fuoriconvegno activities may be found on the website of the meeting[1]. We offer this book to all members of the Italian Statistical Society, to all participants of the conference and to all interested people, in the hope that this will provide them with a good snapshot of the on-going research in this exciting new area of statistical studies. We deeply thank all contributors for having submitted their work to the conference and all the researchers who did an outstanding job in acting as referees accurately and timely. Finally we wish to express our gratitude to the publisher Pearson Italia for all the support received.

Giuseppe Arbia
Stefano Peluso
Alessia Pini
Giulia Rivellini

[1] URL: http://meetings3.sis-statistica.org/index.php/SIS2019/sis2019/schedConf/overview

Predicting and improving smart mobility: a robust model-based approach to the BikeMi BSS

Prevedere e migliorare la mobilità smart: un approccio robusto di classificazione applicato a BikeMi

Andrea Cappozzo, Francesca Greselin and Giancarlo Manzi

Abstract Bike Sharing Systems play a central role in what is identified to be one of the six pillars of a Smart City: smart mobility. Motivated by a freely available dataset, we discuss the employment of two robust model-based classifiers for predicting the occurrence of situations in which a bike station is either empty or full, thus possibly creating demand loss and customer dissatisfaction. Experiments on BikeMi stations located in the central area of Milan are provided to underline the benefits of the proposed methods.

Abstract *I sistemi di Bike Sharing giocano un ruolo centrale nella mobilità sostenibile, uno dei sei pilastri che indentificano una Smart City. Motivati da un set di dati disponibile online, questo lavoro presenta l'utilizzo di due modelli di classificazione robusta per prevedere il manifestarsi di situazioni in cui una bike station sia piena e/o vuota, così creando perdita di domanda ed insoddisfazione nei clienti. Esperimenti di classificazione sulle stazioni BikeMi nel centro di Milano evidenziano l'efficacia dei metodi proposti.*

Key words: Bike Sharing System, Smart Mobility, Impartial Trimming, Robust Classification

1 Motivating problem

The world's population forecast is estimated to reach 9 billions in the upcoming years, with up to 66% of the total humankind living in urbanized areas [6]. The

Andrea Cappozzo • Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: a.cappozzo@campus.unimib.it;francesca.greselin@unimib.it

Giancarlo Manzi
Department of Economics, Management and Quantitative Methods, University of Milan, e-mail: giancarlo.manzi@unimi.it

urban ecosystem devoted to accommodate such a huge proportion of the future population will most likely be a *smart city:* a new metropolitan vision that integrates information and communications technology (ICT) and physical infrastructure, encompassing every municipality aspect: from mobility to architecture, infrastructure and power supply management [7]. Particularly, six pillars identify and assess the concept of "smartness" in such context: economy, people, governance, environment, living and mobility [12].

The present work will focus on sustainable mobility and specifically on the analysis of the BikeMi bike sharing system (BSS) in Milan, as an environmental friendly complement to public and private transports, with the final aim of assessing and possibly improving the service. It is well known that BSS with docking stations users identify finding an available bicycle and a parking slot as the two most critical problems in their biking experience [4]. By employing a robust classification model we try to predict whether and when these problems might occur, identifying some useful insights that may be of use in subsequently planning manual bicycle repositioning.

The rest of the manuscript is organized as follows: in Section 2 the main characteristics of the BikeMi bike sharing system is described; together with the dataset considered in the study. Section 3 details the robust classification method employed in predicting possible future FULL/ EMPTY stations scenarios, with the analysis results presented in Section 4. The paper concludes with a list of proposals for future research direction.

2 The BikeMi BSS

BikeMi was introduced in November 2008 as the first privately managed Italian bike sharing system [9]. Presently, the service encompasses 280 active stations for a total of 4650 available bikes. The dataset considered in this manuscript reports the stations status, in terms of available bikes and free slots, during the period January-August 2015. Records were periodically collected by scraping the BikeMi website: the full dataset is publicly available online [11]. The average weekday profile usage in terms of Normalized Available Bikes (number of bikes / total number of slots in the station) is represented in Figure 1. From the plot two main distinct behaviors are visible: stations that are almost full in the morning and get gradually empty, and stations that follow a mirror pattern. Such a scheme is primarily driven by morning and evening work commuters; it is therefore essential for the BSS success to be able to efficiently cope with this daily bikes demand.

The aim of the present work is to develop a classifier that will help in predicting whether a demand loss might occur. Specifically, since lost demand arises as a consequence of full or empty station, we build a method that predicts how likely are such situations to happen given a set of available information. Time features, past inventory features and meteorological variables are employed in building the classification rule. Given the noisy nature of the dataset at hand (i.e., stations and slots are prone to malfunctions and breakage, undermining the quality of the scrap-

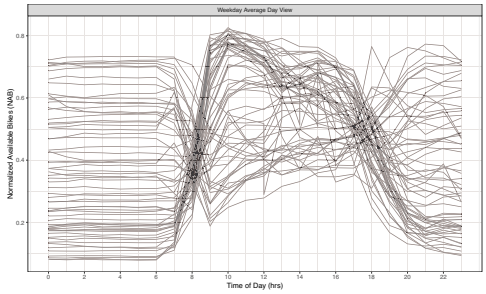

Fig. 1 Average weekday Normalized Available Bikes (number of bikes / total number of slots in the station) for the BikeMi stations in Milan central area.

ing) we propose to employ two robust model-based classifiers for determining the future FULL, EMPTY or NOT PROBLEMATIC status of a particular station. The employed methods are described in the next Section.

3 Robust model-based classifiers

Let $\{(\mathbf{x}_1,\mathbf{l}_1),\ldots,(\mathbf{x}_N,\mathbf{l}_N)\}$ be a complete set of learning observations, where $\mathbf{x}_n$ denotes a p-variate observation and $\mathbf{l}_n$ its associated class label, such that $l_{ng}=1$ if observation n belongs to group g and 0 otherwise, $g=1,\ldots,G$, $n=1,\ldots,N$. In the context of our analysis we set $G=3$, to define the FULL, EMPTY or NOT PROBLEMATIC status of a station in the future time-slot. Likewise, denote the set of unlabelled observations by $\mathbf{y}_m$, $m=1,\ldots,M$ and their associated unknown labels z_{mg}, $g=1\ldots G$ and $m=1,\ldots M$. We construct a procedure for maximizing the *trimmed observed data log-likelihood*:

$$\begin{aligned}\ell_{trim}(\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X},\mathbf{Y},\mathbf{l}) = &\sum_{n=1}^{N}\zeta(\mathbf{x}_n)\sum_{g=1}^{G}\mathrm{l}_{ng}\log\left[\tau_g\phi(\mathbf{x}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right]+\\ &+\sum_{m=1}^{M}\varphi(\mathbf{y}_m)\log\left[\sum_{g=1}^{G}\tau_g\phi(\mathbf{y}_m;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right]\end{aligned} \tag{1}$$

where τ_g is the prior probability of observing class g; $\phi(\cdot;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ is the multivariate normal density with mean vector $\boldsymbol{\mu}_g$ and variance covariance matrix $\boldsymbol{\Sigma}_g$; $\zeta(\cdot)$ and $\varphi(\cdot)$ are 0-1 trimming indicator functions, that express whether observation $\mathbf{x}_n$ and $\mathbf{y}_m$ are trimmed off or not. The *labelled trimming level* α_l, s.t. $\sum_{n=1}^{N}\zeta(\mathbf{x}_n)=\lceil N(1-\alpha_l)\rceil$ and the *unlabelled trimming level* α_u, s.t. $\sum_{m=1}^{M}\varphi(\mathbf{y}_m)=\lceil M(1-\alpha_u)\rceil$ account for possible noisy observations and outliers in both sets.

The aforementioned specification leads to two robust model-based classification approaches: if only the labelled observations are employed for estimating parameters (i.e., only the first line of (1) is considered) we obtain a Robust Eigenvalue

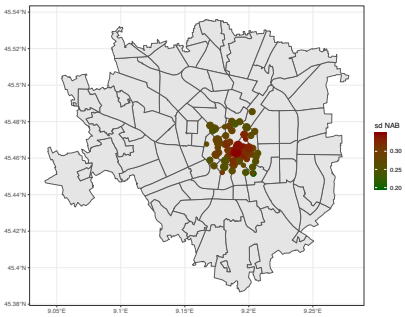

Fig. 2 Location of the BikeMi stations considered in the analysis. Dots size denotes the station total number of slots, the color scaling indicates the standard deviation of normalized bikes availability averaged per weekday.

Decomposition Discriminant Analysis (REDDA); whereas we retrieve a Robust Updating Classification Rule (RUPCLASS) if a semi-supervised approach is favoured. These models are robust generalizations of the techniques developed in [1] and [2], respectively.

Parameters estimation is carried out via a procedure similar to the FastMCD algorithm [8] for the REDDA model, and via the EM algorithm [3] with an appropriate Concentration Step [8] and eigenvalue-ratio restriction [5] enforced at each iteration for RUPCLASS.

4 Classification results

The methodologies described in the previous Section are employed for predicting the station status one hour in the future, thus assessing the need of manual bikes repositioning when FULL and EMPTY situations are forecast. The stations in the analysis are a subset of the ones located in the central area of the city (Bastioni and Centro Storico): a spatial representation is reported in Figure 2.

The considered time-frame is limited to the quarter April-June 2015, in which the last eight days of June are kept out from the learning set and used for assessing the prediction accuracy. The classification results for the models described in Section 3 are reported in Table 1. The classification rates is on average above 0.82 for both REDDA and RUPCLASS, even if the supervised model seems to perform slightly better overall. Particularly, this is more predominant for stations that present a higher turnover, where the unlabelled set provides less useful information about separation between groups [10].

Table 1 Correct classification rates for the BikeMi stations in Milan central area for the last eight days of June 2015 (22-30) employing REDDA and RUPCLASS models.

Station ID	Station Name	REDDA	RUPCLASS
1	Duomo	0.719	0.733
3	Cadorna 1	0.641	0.613
4	Lanza	0.793	0.797
5	Università Cattolica	0.843	0.770
20	Erculea	0.853	0.894
34	Cairoli	0.806	0.788
37	Italia - San Martino	0.931	0.945
43	Festa del Perdono	0.811	0.774
44	Richini	0.912	0.894
45	Cant	0.820	0.811
54	Sant'Eustorgio- P.ta Ticinese	0.853	0.811
60	Edison	0.839	0.816
63	Sant'Ambrogio	0.788	0.806
64	Diaz	0.816	0.811
84	Cadorna 2	0.806	0.848
94	Cadorna 3	0.673	0.668
13	Senato	0.728	0.843
14	San Barnaba H Mangiagalli	0.871	0.912
15	Cantore	0.912	0.912
16	Moscova	0.774	0.779
22	Medaglie D'Oro 1	0.779	0.857
23	Regina Margherita	0.922	0.935
25	Centrale 1	0.687	0.673
27	Porta Venezia	0.797	0.802
30	Crocetta	0.848	0.857
32	Manin - Bastioni	0.880	0.908
46	Porta Nuova	0.876	0.848
55	Cinque Giornate	0.871	0.899
58	Sant'Agostino	0.908	0.889
88	Beatrice d'Este - Cassolo	0.945	0.959
98	San Marco	0.945	0.908
99	Arco della Pace 1 - Bertani	0.903	0.894
103	Arco della Pace 2 - Pagano	0.899	0.889
181	Sempione - Melzi d Eril	0.917	0.922

The present work employs two robust model-based classifiers for detecting possible situations of future demand loss for the BikeMi BSS in Milan. The classification accuracy obtained for a subset of stations in the central area fosters the employment of the described methods. Further research directions will consider the integration of spatial information related to the inventory of the stations closest to the target, and the employment of a cost function for over-penalizing the misclassification of FULL and EMPTY statuses as NOT PROBLEMATIC.

References

1. H. Bensmail and G. Celeux. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91(436):1743–1748, dec 1996.
2. N. Dean, T. B. Murphy, and G. Downey. Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(1):1–14, 2006.
3. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
4. J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI International Joint Conference on Artificial Intelligence*, 2009.
5. S. Ingrassia. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, 13(2):151–166, 2004.
6. S. Mallapuram, N. Ngwum, F. Yuan, C. Lu, and W. Yu. Smart city: The state of the art, datasets, and evaluation platforms. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 447–452. IEEE, may 2017.
7. I.-I. Picioroaga, M. Eremia, and M. Sanduleac. SMART CITY: Definition and Evaluation of Key Performance Indicators. In *2018 International Conference and Exposition on Electrical And Power Engineering (EPE)*, pages 217–222. IEEE, oct 2018.
8. P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, aug 1999.
9. G. Saibene and G. Manzi. Bike usage in public bike-sharing: An analysis of the BikeMi system in Milan. Technical report, 2015.
10. D. Toher, G. Downey, and T. B. Murphy. A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemometrics and Intelligent Laboratory Systems*, 89(2):102–115, nov 2007.
11. A. Trentini. Scraped Data BikeMI 2015. *http://doi.org/10.5281/zenodo.1209270*, mar 2018.
12. J. Zawieska and J. Pieriegud. Smart city as a tool for sustainable mobility and transport decarbonisation. *Transport Policy*, 63:39–50, apr 2018.