Forecasting flood risk exposure using mobile phone traffic flows' data

Rodolfo Metulini[1], Maurizio Carpita[2]

1. Department of Economics and Statistics - University of Salerno
2. Data Methods and Systems Statistical Laboratory - Department of Economics and Management, University of Brescia

Insurance Data Science Conference - Università Cattolica del Sacro Cuore - Milan

June 17^{th}, 2022

The context

- **Floods** are unpredictable natural disasters which cause social and economic impacts on human life. A fast response in evacuating people is a risk management issues (**Mishra et al., 2019**).
- Maps of flood **risk exposure** assume people density **constant** over time, despite this is not the case in urban areas, as crowding is a dynamic process.
- Among ICT and big data sources, **mobile phone** network data allow to obtain **dynamic** information on people's presences (**Metulini & Carpita, 2021**) and movements (**Tettamanti & Varga, 2014; Metulini & Carpita, 2022**), used, e.g.:
  - to develop dynamic exposure to flood risk maps for areas with hydrogeological criticity (**Balistrocchi et al., 2020**).

The project

Project of Lombardy Region, Italy (Infrastrutture e servizi per la Mobilità Sostenibile e Resiliente) - CallHub ID 1180965, bit.ly/2Xh2Nfr

Scientific collaboration with the *Department of Civil, Environmental, Architectural Engineering and Mathematics, UNIBS* (Prof. Roberto Ranzi)

Agenda:

- We have presented at the "European Geosciences Union" (EGU) General Assembly 2022 (with B. Razdar and R. Ranzi) and at the "9th International Conference on Risk Analysis" (ICRA9).
- We will present at the "8th International conference on Time Series and Forecasting" (ITISE 2022) (with P. Zuccolotto, G. De Luca).

Scientific output:

- Carpita, M., Metulini, R. (2021). Modelling the spatio-temporal dynamic of traffic flows with gravity models and mobile phone data, **ASA 2021 Statistics and Information Systems for Policy Evaluation**, Edited by: Bertaccini, B.; Fabbris, L.; Petrucci, A.
- Metulini, R., Carpita, M. (2021). A Spatio-Temporal Indicator for City Users based on Mobile Phone Signals and Administrative Data - **Social Indicator Research**, 156, 761–781.
- Balistrocchi, M., Metulini, R., Carpita, M., and Ranzi, R. (2020). Dynamic maps of people exposure to floods based on mobile phone data. **Natural Hazards and Earth System Sciences**, 20, 3485–3500.
- Metulini, R., Carpita, M., Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making; (Submitted to **Annals of Operations Research**)

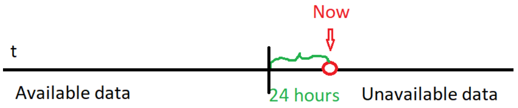
Mobile phone traffic flows data

- Provided by Olivetti and FasterNet for the MoSoRe Project.
- 235×235 OD matrices (where each row/column is an ACE in the province of Brescia), available at hour intervals from September, 1^{st} 2020 to August, 31^{th} 2021.
- **Interpretation**: amount of traffic flows from a specific ACE to another specific ACE in that hour interval, where flows from a specific ACE to itself (also called "internal" flows) are displayed in the main diagonal of the OD matrix.
    - In this work, we just consider human SIM[1] (M2M machine SIM are excluded) in order to avoid duplicates (e.g. a person who has both the mobile phone and the black box in his/her car).
    - Data present the limitation that locations of the SIM are retrieved every 5 minutes. This can lead to an underestimation of the amount of flows.

[1] Both Italian and foreigner (in roaming) SIM are counted.

Our specific goal

- For technical reasons data are provided with some hours of delay.

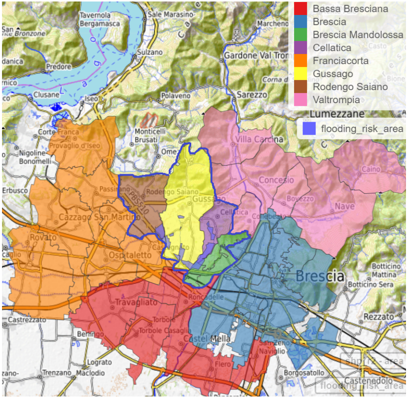


- From a risk management perspective, this fact prevents decision makers from knowing the current amount of people to warn or to evacuate.
- From a **statistical** point of view, our aim is to develop a **time-series** model to **forecast** the current flow of people which meets the following requirements:
  1. it presents a good predicting performance;
  2. its functioning does not rely on recent (latest 24 hours) data.
- Using data on mobile phone **Origin-Destination flows** we show an application the case study of *Mandolossa*[2].

[2] A critical zone with flood episodes in the north-west of Brescia

Area of interest

- The ACEs intersecting the flooding-risk area are Gussago, Cellatica, Rodengo Saiano and Brescia Mandolossa.
- We have identified other 38 ACEs that present a strong flows' intensity from/to the Mandolossa (in turn aggregated in four macro areas: Bassa bresciana, Brescia, Valtrompia and Franciacorta).



Preliminary evidences

- We studied the temporal dynamic of traffic flows by jointly analyzing the three following time series[3,4]:
  1. Flows to Cellatica from all other 38 neighboring ACEs ("inflows");
  2. flows from Cellatica to all other 38 neighboring ACEs ("outflows");
  3. flows from Cellatica to Cellatica itself ("internal flows").
- A strong similarity among the different time series, likely due to the topographical characterization of urbanized areas, motivates us to model such flows as belonging to **dependent processes**.
- ACF and PACF highlight strong **daily seasonal patterns**, which is very similar when considering inflows, outflows or internal flows.
- An additive decomposition of the time series in `trend`, daily (`season_24`) and weekly (`season_168`) seasonality, obtained using STL with LOESS (Cleveland et al., 1990) highlights the presence of a strong **weekly seasonal pattern**.

[3] We analyse Cellatica because it is the ACE with the largest intersection with the flooding risk map.

[4] We consider both inflows and outflows and internal flows because, to correctly quantify street crowding, it is also necessary to consider those moving within the borders of the considered area.

Model specification

- A VAR model with eXogenous variables (**VARX**, **Tsay, 2014**, ch. 6) as it follows:

$$\mathbf{y}_t = \boldsymbol{\nu} + \underbrace{\sum_{h_d=1}^{p_d} \mathbf{A}_{h_d} \mathbf{y}_{t-h_d}}_{\text{daily AR term}} + \underbrace{\sum_{h_w=1}^{p_w} \mathbf{A}_{h_w} \mathbf{y}_{t-h_w}}_{\text{weekly AR term}} + \underbrace{\mathbf{B}\mathbf{x}_t}_{\text{exogenous / DHR term}} + \epsilon_t \quad (1)$$

  - p_d is the daily AR order (h_d=1 $\rightarrow$ 24 hours = 1 day)
  - p_w is the weekly AR order (h_w=1 $\rightarrow$ 24*7 hours = 1 week)

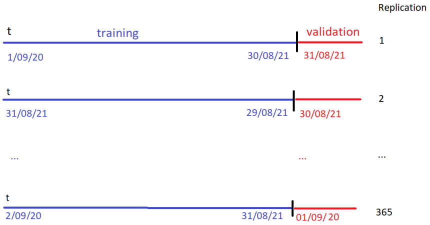
- For the exogenous part $\mathbf{B}\mathbf{x}_t$ we employ a **DHR** structure (**Hyndman & Athanasopoulos, 2021**) which we define as a sum of daily and weekly Fourier bases. For a given VAR equation and a given t:

$$\underbrace{\sum_{k_d=1}^{K_d} [\alpha_{k_d} s_{k_d}(t) + \gamma_{k_d} c_{k_d}(t)]}_{\text{daily pattern}} + \underbrace{\sum_{k_w=1}^{K_w} [\alpha_{k_w} s_{k_w}(t) + \gamma_{k_w} c_{k_w}(t)]}_{\text{weekly pattern}} \quad (2)$$

  - $s_{k_d}(t) = sin(\frac{2\pi kt}{m_d})$, $c_{k_d}(t) = cos(\frac{2\pi kt}{m_d})$, $s_{k_w}(t) = sin(\frac{2\pi kt}{m_w})$, $c_{k_w}(t) = cos(\frac{2\pi kt}{m_w})$, $m_d = 24$. $m_w = 168$.

Performance evaluation

- We generate different training and validation sets (one for each different day of the year) by properly reordering the dataset.



- A two step evaluation is proposed, where:
  1. **MAPE** is used to **detect** and remove **outlier days**

$$MAPE = 100 \times \frac{1}{n_v} \sum_{t=1}^{n_v} \frac{\mid Y_t - \hat{Y}_t \mid}{Y_t} \tag{3}$$

  2. By previously assigning observed and predicted values of the validation set to categories using distribution quintiles, **HR** is adopted to **evaluate the performance** in assign days to categories.

$$HR = \frac{1}{n_v} \sum_{t=1}^{n_v} I(\; Y_t \text{ and } \hat{Y}_t \text{ belong to same category}) \tag{4}$$

VARX model calibration

- The chosen estimated model is a VARX(p_d=3, p_w=4) model with a DHR($K_d = 7, K_w = 4$) term among exogenous variables[5]:
  - Optimal number of Fourier bases are determined with the method based on minimizing AIC, suggested in previous studies for the case of single (**Hyndman & Athanasopoulos, 2021**) and multiple (**Metulini & Carpita, 2022**, sec. 3) patterns.
  - For the identification of the AR order a two-step approach based on AIC and the "elbow" method has been applied.
- The diagnostic on the model residuals displays the presence of significant autocorrelation at some orders and a leptokurtic distribution with heavy tails. This is the price to pay for does not including traditional lag terms in the model.

[5] Performed in R based on `VARX` and `VARXpred` functions in `MTS` package.

Forecasting evaluation

As an example, the forecast for 4 sample days of February 2021.

Schematically we proceed as it follows:

1. we analyze by means of the MAPE the forecasting accuracy of our method on the original dataset. In doing so, we also detect outlier days.
   - For the days detected as outliers we have replaced the values of the inflows, the outflows and the internal flows with the values of the same weekday of the previous week.
2. in a second step, on a the dataset where outlier days have been appropriately replaced we analyse by means of the HR the performance of our method in correctly **detecting days with high, moderate or low levels of flows**.

Conclusions

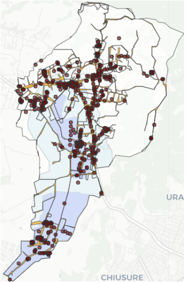
- Our novel methodological strategy (modelling + forecasting), based on VARX + DHR and on a two step validation with MAPE and the HR indices:
  1. presents a good performance in predicting whether current traffic flows in flooding risk areas is "very high", "high", "moderate", "low", "very low".
  2. permits to detect those year's days for which the traffic flows' forecast based on our method is not good enough.
- Public decision-makers can take advantage of our statical method based on mobile phone data that can be used for real-time predictive purposes to deal with future (flood) emergency situations.

Discussion & Future developments

- The poor prediction on anomalous days might be due the limitation given by not including the previous hours lag terms in the model.
- This limitation may turn out to an advantage. In fact, we prove that a fairly good predicting performance on traffic flows may be obtained even if previous hours' information is not accounted for.

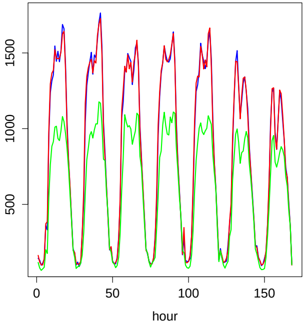
- Ongoing research for the MoSoRe project are devoted to:

1. apply the method to other flooding risk areas.
2. rescale O-D flows data in flooding risk areas with the help of Minimization Drive Test (MDT) technology data. (**See Figure** →)
3. Allow leptokurtic distribution of residuals to be modelled with VARX.

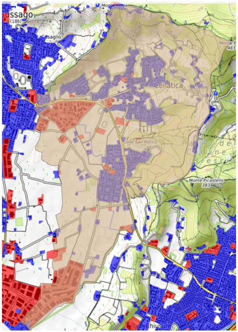


References

1. Balistrocchi, M., Metulini, R., Carpita, M., & Ranzi, R. (2020). Dynamic maps of human exposure to floods based on mobile phone data. Natural Hazards and Earth System Sciences, 20(12), 3485-3500.
2. R. B. Cleveland, W. S. Cleveland, J. E. McRae, I. J. Terpenning, STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1) (1990) 3–33.
3. Metulini, R., Carpita, M. (2022). Forecasting Traffic Flows with Complex Seasonality using Mobile Phone Data - IES 22, INNOVATION AND SOCIETY 5.0: STATISTICAL AND ECONOMIC METHODOLOGIES FOR QUALITY ASSESSMENT
4. Hyndman, R. J., Athanasopoulos, G. (2021). Forecasting: principles and practice. 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3.
5. Metulini, R., & Carpita, M. (2021). A spatio-temporal indicator for city users based on mobile phone signals and administrative data. Social Indicators Research, 156(2), 761-781.
6. D. Mishra, S. Kumar & E. Hassini, Current trends in disaster management simulation modelling research. Annals of Operations Research, 283(1) (2019) 1387-1411.
7. Tettamanti, T., Varga, I. (2014). Mobile phone location area based traffic flow estimation in urban road traffic. Columbia International Publishing, Advances in Civil and Environmental Engineering, 1(1): 1-15.
8. R.S. Tsay, Multivariate Time Series Analysis with R and Financial Applications (2014) John Wiley & Sons. Hoboken, NJ.

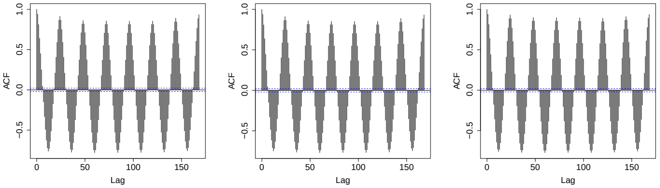


(a) Outflows (in red), inflows (blue), internal flows (green). Hourly data from 22th to 28th of February, 2021.

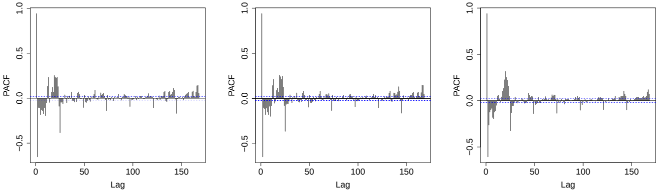


(b) Map of the ACE of Cellatica, by destination of use. Residential areas are depicted in blue, industrial areas in red. Source: Destinazione d'Uso dei Suoli Agricoli e Forestali (DUSAF) 6.0 2018.

Back to slide



(a) ACF.



(b) PACF.

Figure: ACF and PACF for Inflows (left), outflows (middle) and internal flows (right). Time lags up to one week (168 hours).

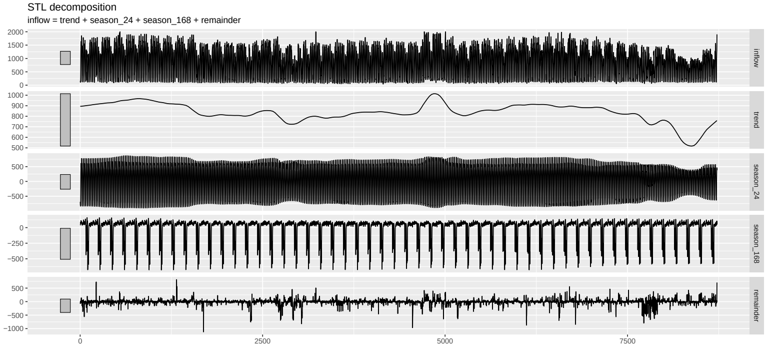


Figure: STL with LOESS: trend, daily, weekly patterns and a remainder component. Data from September 1^{st}, 2020 to August 31^{st}, 2021. For simplicity, just inflows are reported.

Back to slide

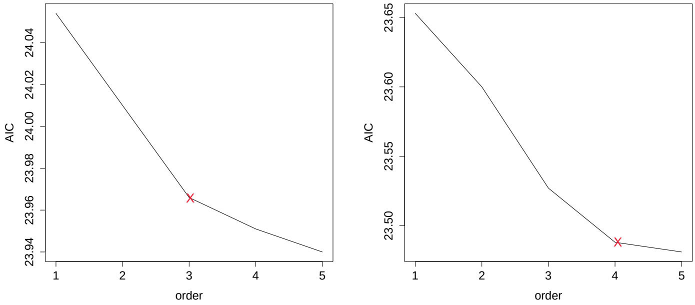


Figure: AIC of the VARX model with Fourier bases ($K_d = 7$, $K_w = 4$), month, weekday dummies and $p_w = 0$, for $p_d = 1, 2, ..., 5$ (left). AIC of the VARX model with Fourier bases ($K_d = 7$, $K_w = 4$), month, weekday dummies and $p_d = 3$, for $p_w = 1, 2, ..., 5$ (right).

Back to slide

Table: Results of the VARX($3_d, 4_w$) with a DHR($7_d, 4_w$) component.

endogenous variable	**inflow (s.e.)**	**outflow (s.e.)**	**internal flow (s.e.)**
inflow_AR(1)_day	0.192 (0.056)	0.063 (0.057)	-0.009 (0.037)
inflow_AR(2)_day	-0.051 (0.057)	-0.096 (0.058)	-0.036 (0.036)
inflow_AR(3)_day	0.028 (0.057)	-0.003 (0.057)	-0.039 (0.036)
outflow_AR(1)_day	0.109 (0.056)	0.230 (0.056)	0.035 (0.036)
outflow_AR(2)_day	-0.072 (0.057)	-0.032 (0.057)	-0.058 (0.036)
outflow_AR(3)_day	-0.006 (0.056)	0.024 (0.056)	-0.006 (0.035)
internal_flow_AR(1)_day	0.035 (0.032)	0.042 (0.032)	0.258 (0.021)
internal_flow_AR(2)_day	0.160 (0.033)	0.170 (0.033)	0.165 (0.020)
internal_flow_AR(3)_day	0.133 (0.032)	0.133 (0.033)	0.176 (0.020)
inflow_AR(1)_week	0.294 (0.059)	0.136 (0.060)	0.099 (0.037)
inflow_AR(2)_week	-0.017 (0.058)	-0.133 (0.059)	-0.016 (0.037)
inflow_AR(3)_week	0.131 (0.059)	0.048 (0.060)	0.039 (0.037)
inflow_AR(4)_week	0.168 (0.059)	0.084 (0.059)	0.073 (0.037)
outflow_AR(1)_week	0.012 (0.059)	0.171 (0.059)	-0.044 (0.037)
outflow_AR(2)_week	0.070 (0.058)	0.186 (0.059)	0.008 (0.037)
outflow_AR(3)_week	0.120 (0.058)	0.207 (0.059)	0.050 (0.037)
outflow_AR(4)_week	0.077 (0.059)	0.160 (0.060)	0.022 (0.037)
internal_flow_AR(1)_week	-0.067 (0.032)	-0.069 (0.033)	0.145 (0.021)
internal_flow_AR(2)_week	-0.058 (0.031)	-0.061 (0.032)	0.046 (0.020)
internal_flow_AR(3)_week	-0.217 (0.032)	-0.221 (0.031)	-0.044 (0.020)
internal_flow_AR(4)_week	-0.174 (0.032)	-0.168 (0.032)	-0.041 (0.020)
exogenous variable	**outflow (s.e.)**	**inflow (s.e.)**	**internal flow (s.e.)**
sin_day_1	-45.439 (7.928)	-37.692 (7.990)	-47.087 (5.015)
cos_day_1	-25.347 (7.432)	-29.108 (7.490)	-27.341 (4.701)
sin_day_2	-9.508 (2.976)	-11.809 (3.000)	-14.312 (1.883)
cos_day_2	5.004 (2.232)	3.044 (2.249)	5.749 (1.412)
sin_day_3	4.342 (2.579)	5.511 (2.600)	4.179 (1.632)
cos_day_3	-11.551 (3.188)	-4.073 (3.213)	-7.625 (2.016)
sin_day_4	-0.254 (2.086)	-2.125 (2.103)	-3.108 (1.320)
cos_day_4	1.540 (2.473)	-2.723 (2.493)	-3.912 (1.565)
sin_day_5	0.166 (2.086)	0.348 (2.103)	1.962 (1.320)
cos_day_5	0.327 (1.956)	-0.486 (1.971)	-0.159 (1.237)
sin_day_6	1.029 (1.941)	0.662 (1.956)	0.839 (1.228)
cos_day_6	-1.805 (1.969)	-0.736 (1.985)	-1.310 (1.246)
sin_day_7	-1.472 (2.023)	0.425 (2.039)	-0.010 (1.280)
cos_day_7	-1.026 (1.969)	-1.361 (1.985)	-1.143 (1.246)
sin_week_1	41.318 (2.358)	41.137 (2.377)	22.975 (1.492)
cos_week_1	5.778 (2.347)	6.137 (2.366)	5.785 (1.485)
sin_week_2	-35.458 (2.324)	-35.469 (2.342)	-22.282 (1.470)
cos_week_2	10.698 (2.177)	9.993 (2.195)	2.815 (1.377)
sin_week_3	10.185 (1.978)	10.820 (1.994)	6.417 (1.251)
cos_week_3	-23.977 (2.113)	-23.635 (2.129)	-10.739 (1.337)
sin_week_4	16.133 (1.985)	15.494 (2.000)	9.308 (1.256)
cos_week_4	18.184 (2.069)	18.348 (2.086)	7.757 (1.309)
month (ref. January): February	43.963 (6.942)	44.745 (6.997)	3.117 (4.392)
March	56.449 (6.775)	56.982 (6.829)	23.425 (4.286)
April	7.840 (6.807)	8.286 (6.861)	-0.302 (4.306)
May	43.830 (7.116)	44.526 (7.172)	7.260 (4.501)
June	65.148 (7.097)	66.822 (7.153)	21.094 (4.490)
July	68.355 (7.234)	69.637 (7.291)	9.575 (4.576)
August	-32.542 (7.210)	-33.130 (7.267)	-47.680 (4.561)
September	42.524 (7.544)	44.453 (7.604)	11.643 (4.773)
October	102.004 (7.127)	103.378 (7.183)	38.790 (4.508)
November	54.185 (6.947)	54.458 (7.002)	31.033 (4.394)
December	71.030 (6.794)	71.378 (6.847)	27.408 (4.298)
weekday (ref. Monday): Tuesday	63.740 (26.874)	68.238 (27.087)	31.087 (17.001)
Wednesday	11.398 (26.888)	17.867 (27.101)	-13.700 (17.009)
Thursday	11.876 (26.920)	23.314 (27.133)	-10.966 (17.029)
Friday	9.321 (26.875)	27.938 (27.088)	-27.546 (17.090)
Saturday	-64.434 (26.842)	-46.322 (27.055)	-59.964 (16.980)
Sunday	-30.698 (26.858)	-22.364 (27.071)	-20.947 (16.990)
intercept	19.612 (13.058)	19.244 (13.161)	61.881 (8.260)
residual correlation matrix	outflow	inflow	internal flow
outflow	1	0.983	0.858
inflow	0.983	1	0.856
internal flow	0.858	0.856	1
information criteria	AIC: 23.488	BIC: 23.636	

Back to slide

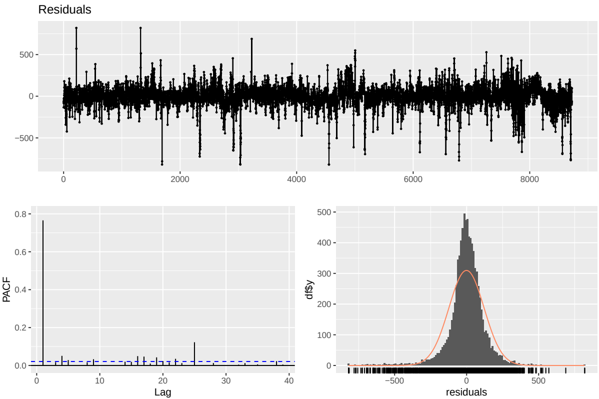


Figure: Estimated residuals (top), PACF with 95% confidence bounds for strict white noise (bottom left), histogram of the empirical distribution (bottom right). For simplicity, just diagnostic for inflows is reported.

Back to slide

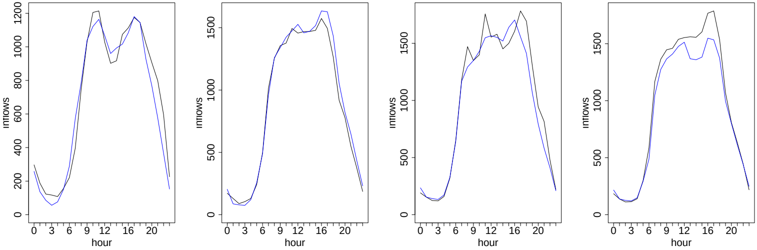


Figure: Plot of observed (black) versus predicted (colored) traffic flow in Cellatica. Validation days (from left to right): February, 15th (Monday), February, 17th (Wednesday), February, 19th (Friday), February, 21st (Sunday) 2021. For simplicity, just inflows are reported.

Back to slide

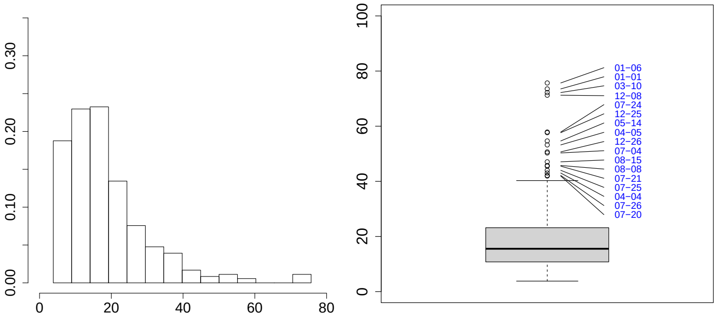


Figure: Histogram (left) and box plot (right) of the distribution of the MAPE computed based on all validation sets. For simplicity, just inflows are reported.

Back to slide

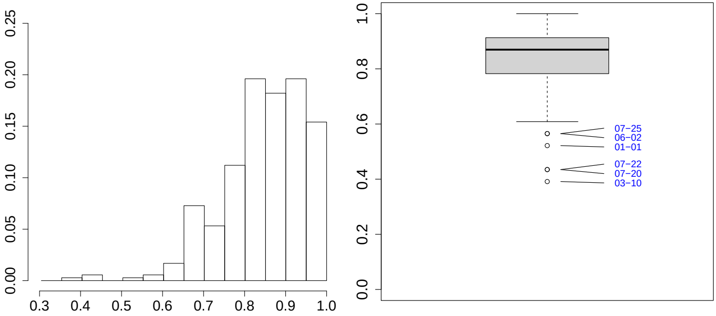


Figure: Histogram of the HR based on all validation sets, where 1^{st} step outliers are replaced (left), box plot of the HR based on all validation sets, where outliers found in the 1^{st} step of detection are already replaced. 2^{nd} step outliers labeled in blu (right). For simplicity, just inflows are reported.

Back to slide