

# Short-term forecasts on time series for tourism in Lombardy

Andrea Marletta<sup>a</sup>, Roberta Rossi<sup>b</sup>, Elena Diceglie<sup>b</sup>

<sup>a</sup>University of Milano-Bicocca, Department of Economics, Management and Statistics

<sup>b</sup>Polis-Lombardia

## 1. Introduction

Data from official statistics are often available only some months after their collection. For this reason, above all in the public sector, there are a lot of requests about predictions or provisional data for the statistical phenomena in order to be distributed as soon as possible. In the last years, after the pandemic emergency due to Covid-19, the curiosity of public stakeholders about the existence of an economic recover after the restrictions of 2020 (and partially of 2021) has increased this request to have very quickly official data.

The aim of this paper is to try to accelerate this process proposing short-term predictions in time series as temporary substitutes waiting for official data. The context of this work is in the tourism sector, one of the most damaged economic sectors by the limitations due to Covid-19. Many contributions are already present in literature about the strategy and the estimation for the recovery of the travel sector after the pandemic emergency (Fotiadis et al., 2021; Yeh, 2021). In this context, the main objective of this work is to verify the presence of a full or partial recover of tourists in provinces of Lombardy using short-term predictions for 2022. This issue has also been treated by Provenzano and Volo (2022). This contribution is the result of a collaboration with PoliS-Lombardia, a public institution of Regione Lombardia. It is included in the list of institutional units belonging to the public sector published by Istat.

PoliS-Lombardia has been instituted in 2018 and it is the regional institute for the support to the policies of Lombardy. Its mission is the implementation and the evaluation of the policies in Lombardy. The main functions of PoliS-Lombardia are: support to the integrated policies of education and labour coherently with fixed objectives by the administration; studies and research projects related to the institutional, local, economic and social processes; management of the regional statistical function in collaboration with ISTAT; management and coordination of the regional observatories; education of the regional employees. Given this scopes, it represents a very important stakeholders in the field of data management in Lombardy involved in a large amount of data, as for example in the tourism sector.

In this paper, using a short-term forecasts approach, some preliminary results will be presented for detecting a recovery in the travel sector for 2022 using the total number of presences in Lombard provinces. These short-term predictions will be obtained using a very well-known methodology in time-series literature, such as the ARIMA (Auto-Regressive Integrated Moving Average) models (Box et al., 2015; Hamilton, 2020; Wei, 2006). In these models, an exogenous variable representing the working positions in the food services and hospitality industry has been added supposing an high correlation between the two phenomena.

## 2. Methodological tools

Data from official sources on nights spent in an accommodation for tourists in Lombardy are available until 2021. These data on travel flows for 2020 and 2021 registered a clear downfall because of restrictions related to Covid-19.

A time-series procedure has been applied to obtain a forecast estimate for 2022 using an ARIMA model with the addition of an exogenous variable.

The ARIMA (Auto-Regressive Integrated Moving Average) models have been introduced as mixed models composed by an Auto-Regressive (AR) part in which the single observation depends on the lagged values of the time series, a Moving Average (MA) part in which the same observation depends on the lagged values of the errors and, if necessary, an Integrated (I) part considering the original time series in differences according an integration order (Wei, 2006).

They could be represented as:

$$\phi_p(B)(1 - B)^d Z_t = \theta_q(B)a_t$$

where  $\phi_p(B)$  represents the AR part,  $(1 - B)^d Z_t$  the I part and  $\theta_q(B)a_t$ .

The hypothesis at the basis of the model is that a punctual estimate of the travel flows could be obtained using an auxiliary variable explaining the number of employees in the food services and hospitality industry. Statistically speaking, this means to introduce ARIMAX models, that is to say, ARIMA models with an exogenous variable with the following notation:

$$\phi_p(B)(1 - B)^d Z_t = \theta_q(B)a_t + \beta_i x_i$$

where  $\beta_i x_i$  is the X part of the model. This auxiliary variable is represented as the difference between the number of starting work contracts and the contract terminations. These data are available thanks to the Informative system of mandatory communications provided by the Italian Minister of Labour. The availability of this information is daily guaranteed at level of single municipality but for the purpose of this paper, data have been aggregated at province level.

The short-term predictions obtained for 2022 have been used to verify the presence of a recovery respect to the pandemic emergency of Covid-19 using a double growth rate. A first growth rate has been computed comparing the number of estimated tourists respect to the 2021 measuring the existence of a rebound after the restrictions. A second growth rate measured the estimates for 2021 respect to the presences of 2019 to monitor the trends in Lombardy compared to the before Covid-19 period.

Data used for the prediction model refers to the total number of travel presences expressed in terms of nights in accommodation from 2017 to 2021. About the auxiliary variable, data refers to the balance expressed as the difference between the activations and the terminations of the job contracts until March 2022. All the elaborations have been computed using R following the approach proposed by Hyndman and Athanasopoulos (2018).

The approach to obtain this short-term forecasts is based on a two-step procedure: firstly, data about employees are predicted for the interval from April to December 2022; secondly, predictions for tourism presences are obtained for the entire 2022.

The time series of the COB (Comunicazioni OBbligatorie) related to activations and terminations of job contracts for the food services and hospitality industry is updated until March 2022. Since PoliS-Lombardia is interested in predicting the entire year 2022, before applying the ARIMAX model, the values for this variable for the remaining months of 2022 have been obtained using a well-known approach choosing the best model among different time-series predictors as ARIMA models and ETS (Error, Trend, Seasonality) models. The model was selected minimizing the Mean Squared Error (MSE).

Once obtained the extended time series on the balance of the job contracts, this can be used as auxiliary variable for predicting the 2022 observations for the travel indicator using an ARIMAX model.

### 3. Application and results

Data source used for the prediction about the total number of travel presences from 2017 to 2021 has been achieved from 2 different surveys. From 2017 to 2020, data are the official statistics released by Istat, for 2021 data are from Istat but they are obtained in a different way and they are still provisional.

The integration of data using provisional information about 2021 has been necessary to obtain plausible forecasts. Without this operation, data about 2020 would have deeply conditioned the predictions in a negative trend. The 2020 data have been influenced by the restrictions due to the pandemic emergency due to Covid-19. Since the Lombard tourism is characterized by seasonality (above all in the mountain provinces), the predictions take into account this aspect underlining different trends for each territory.

Data about start and end of the job contracts are sourced to the Informative system of mandatory communications provided by the Italian Minister of Labour. Since they are computed as a difference, they could assume positive and negative values. They are only referred to positions in the food services and hospitality industry. In particular, the hypothesis behind this choice is that an increase of the balance (and therefore of the activations) of the employees in this sector is a symptom of a higher request due to an increase of the travel presences. If these two series are highly correlated, it makes sense to use this variable as exogenous in explaining the travel indicator.

All data are available monthly and from a geographic point of view, they referred to Lombard provinces. In Lombardy, 12 provinces are present, they are: Bergamo, Brescia, Como, Cremona, Lecco, Lodi, Mantova, Milan, Monza-Brianza, Pavia, Sondrio, Varese. In Figure 1, a time series plot with real (in black) and predicted values (in blue) is displayed as an example for Bergamo and Varese provinces.

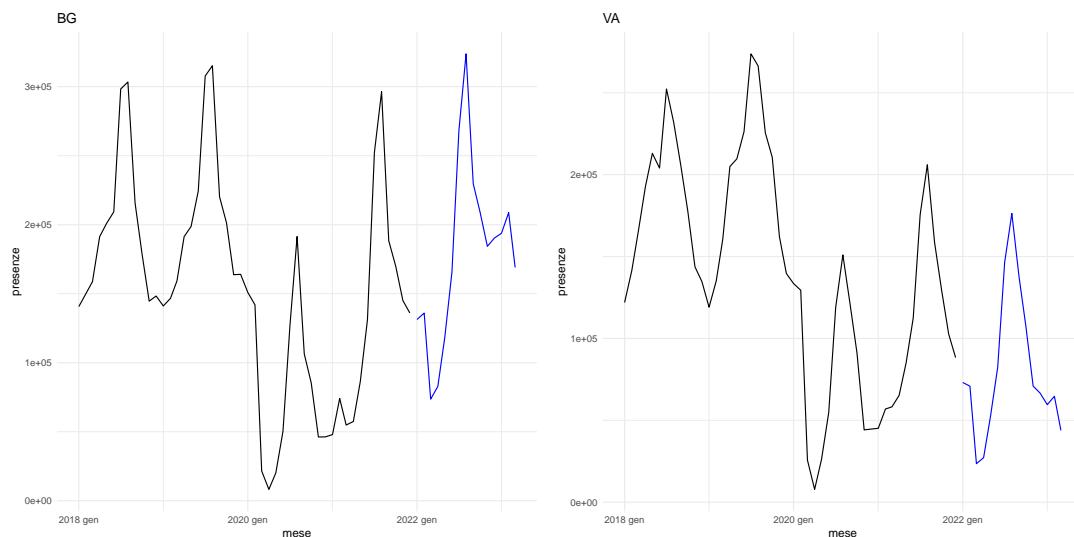


Figure 1: Time series plot for total presences for Bergamo and Varese provinces

As mentioned in the previous section, the research question of the paper is two-fold: firstly, to evaluate the plausible upswing for predicted values for 2022 respect to 2021 and secondly, to compare this predictions with the pre-Covid19 period such as 2019. The answer to this research question could be obtained using two simple growth rates:

$$t_1 = \frac{\text{predicted presences}_{2022}}{\text{official presences}_{2021}} * 100$$

$$t_2 = \frac{\text{predicted presences}_{2022}}{\text{official presences}_{2019}} * 100$$

The results of the model predict a substantial recovery of the Lombard tourism compared to 2021 for almost the 12 provinces with  $t_1$  growth rate higher than 40% in Como, Cremona and Sondrio provinces. Complete results for  $t_1$  are displayed in Figure 2.

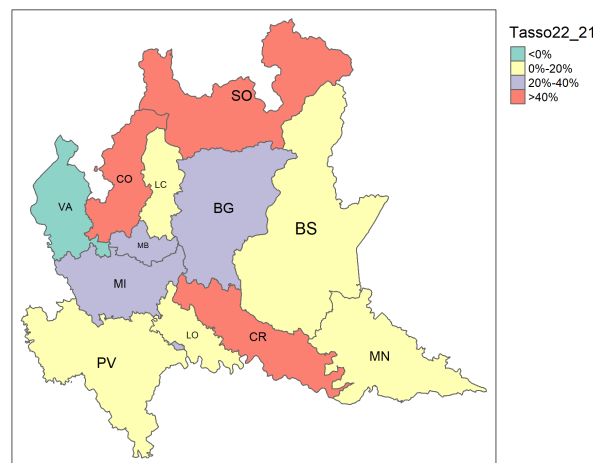


Figure 2: Growth rate per total presences in Lombardy provinces between 2022 and 2021

From the map, it is possible to note that  $t_1$  is positive for all provinces except than Varese. The highest values for  $t_1$  is for Sondrio, where the model estimated a doubling of the presences, but this is due to the fact that Sondrio is a mountain province in which 2021 has been strongly conditioned by the limitations in the winter season. Bergamo, Milan and Monza-Brianza have a growth rate between 20% and 40%. For other provinces it has been registered a moderate growth.

On the other hand, there is not a complete recovery respect to the pre-Covid19 period. Only 4 provinces have positive values for  $t_2$ : Como, Cremona, Monza-Brianza and Sondrio. Complete results for  $t_1$  are displayed in Figure 3.

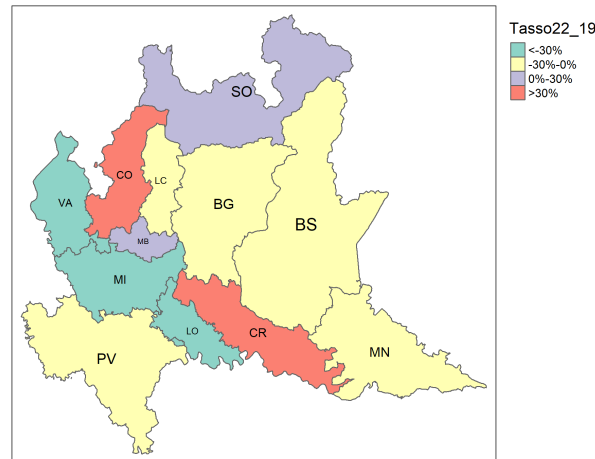


Figure 3: Growth rate per total presences in Lombardy provinces between 2022 and 2019

All the other provinces of the East Lombardy registered a light decline respect to 2019, ma for some provinces as Brescia and Lecco, this decrease is only about 3%, hoping for a complete recovery in 2023. Negative growth rates more stressed are obtained for Lodi, Milan and Varese where the predicted values for presences are still 30% less than 2019, symptom of a slowest recovery.

#### 4. Summary and conclusions

The aim of this paper was to obtain short-term predictions about total presences in tourism sector in 2022 for Lombard provinces using an ARIMAX model considering data from labour market as auxiliary variable. This variable has been used hypothesizing a high correlation between the activations of contracts in food and hospitality sector and the increase of the travel presences. Preliminary results showed an evident upswing respect to 2021 and a partial recovery respect to 2019 for the majority of Lombard provinces. In particular, Sondrio is the province with the highest growth rates and Varese the province with the lowest growth rates.

Future works could focus the attention on other exogenous variables to add in the ARIMAX model hypothesizing other possible influences on the phenomena of the Lombard tourism. The same model could be also replicated for single municipalities or particular industrial districts. Finally, from a methodological point of view, some other prediction techniques could be added as comparison like for example the VAR (Vector Auto-Regressive) models and the relation between presences and workers could be enhanced through a co-integration analysis.

## References

- Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
- Fotiadis, A., Polyzos, S., Huan, T. C. T. (2021). The good, the bad and the ugly on COVID-19 tourism recovery. *Annals of tourism research*, 87, 103-117.
- Hamilton, J. D. (2020). Time series analysis. Princeton university press.
- Hyndman, R. J., Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
- Provenzano, D., Volo, S. (2022). Tourism recovery amid COVID-19: The case of Lombardy, Italy. *Tourism Economics*, 28(1), 110-130.
- Wei, W. W. (2006). Time series analysis: univariate and multivariate. Methods. Boston, MA: Pearson Addison Wesley.
- Yeh, S. S. (2021). Tourism recovery strategy against COVID-19 pandemic. *Tourism Recreation Research*, 46(2), 188-194.