TOURISTS' PRESENCE FLASH ESTIMATION BY MNO DATA

Cristina Faricelli, Alessandro Piovani, Tiziana Tuoto

Abstract. The use of big data to anticipate the release of official statistics is attracting growing attention and interest within the scientific community. This paper presents a methodological contribution to flash estimation of overnight stays in Italian municipalities using administrative and mobile network operator (MNO) data. Three forecasting strategies: Simplistic, Quasi-Transfer Learning (QTL), and Augmented Learning (AL) are compared across different models (Linear, Random Forest, and Ensemble). The application focuses on Emilia-Romagna municipalities, evaluating predictive accuracy through repeated cross-validation. Results show that both QTL and AL outperform traditional methods, especially in terms of mean absolute percentage error (MAPE) and confidence interval stability.

1. Introduction

In many problems, such as territorial forecasting, the availability of data is often asynchronous: some observations are reported earlier than others due to administrative and logistic factors. At the same time, highly informative proxy variables are often available, providing valuable signals for anticipating missing information. This motivates the use of forecasting techniques to produce flash estimates.

The asynchronous availability of the response variable and the presence of good proxies create an opportunity to improve standard forecasting approaches. To explore this idea, we evaluate two advanced methodologies - Quasi-Transfer Learning and Augmented Learning - which aim to refine predictions by exploiting structural patterns in the available data. Specifically, we assess their effectiveness in forecasting the number of nights spent by tourists in a given municipality, using statistical register data and mobile network operator (MNO) proxies. We compare these approaches against the traditional forecasting method.

2. Data description

In the field of tourism statistics, a crucial indicator is the total number of nights spent by tourists. In Italy this indicator is provided for each municipality (local administrative unit) on a monthly basis.

Local offices collect the data provided by touristic accommodations and forward them to the National Statistical Institute, typically within two months ($t+2$) from the reference month (t), to ensure no accommodations remain out of the enumeration. However, a certain percentage of municipalities can provide this information earlier, sometimes even within the first ten days of the reference month. It is worth noting that the set of these early response units may vary each month due to several contingency factors.

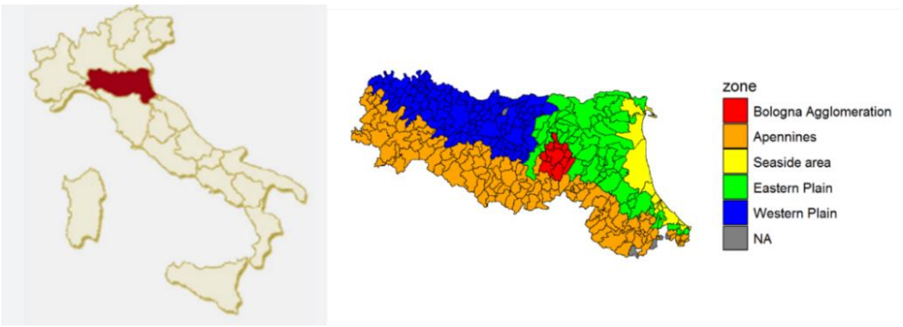
The objective of our study is to investigate whether, by leveraging register data from responding municipalities and the number of overnight stays derived from MNO data, it is possible to provide a flash estimate for municipalities that do not respond promptly.

For this study monthly time series for official overnight stays at municipality level, from October 2023 back to January 2022 are available while MNO data cover the period from August 2022 to October 2023. The data pertain to the Emilia-Romagna which comprises 330 municipalities, grouped in 9 provinces.

The MNO data have been processed according to multi-MNO algorithms; they have also been subject to some adjustments to "reproduce" Istat known totals. The MNO data come from a single operator, and the number of nights is counted based on a simple shared algorithm. Among other aspects, the algorithm includes a rescaling process to account for other operators and factors beyond our control, as well as masking for values at risk of re-identification.

The Emilia Romagna regional territory can be divided into five areas as it is shown in Figure 1: Western Plain, Eastern Plain, Bologna Agglomeration, Apennines and Seaside area.

Figure 1 – *Emilia Romagna region with geographic classification zones.*



From a tourism perspective, the region is primarily characterized by a high volume of seaside tourism along the coast and urban tourism in the provincial capitals, while nature-based tourism in the Apennines and the Po Delta plays a secondary role, with clear and marked seasonal differences. Table 1 shows the different volumes of tourism in the various geographical zones and the number of municipalities in each zone according to register and MNO data.

Table 1 – *Distribution of municipalities in each geographical zone and percentage of nights spent by tourists, according to Register Data and MNO Data.*

Zone	Municipalities	Register Data (Nights/Month %)	MNO Data (Nights/Month %)
Bologna	12	14.1%	20.8%
Apennines	120	0.5%	1.6%
Seaside Area	14	82.6%	70.4%
Eastern Plain	68	1.3%	3.7%
Western Plain	112	1.4%	3.5%

In addition to data on tourist overnight stays, we also have access to auxiliary administrative data for each municipalities, such as the resident population, the surface area, the geographic classification zones and degree of urbanization.

3. Methodological approaches

3.1. *Simplistic approach*

Given a dependent response variable y and some features x which become available before the response, forecasting techniques are used to anticipate *y* based on *x*. The most straightforward approach to this task is what we will call the Simplistic Model. This consists of training a predictor $\mu(x_j, s_t)$ on a training set *s*, where both *y* and *x* are available, and then applying it to a target set *r*, where *x* is available but some values of *y* are still missing. Formally, we estimate:

$$\hat{y}_j = \mu(x_j, s_t), \forall j \notin s_t$$

where s_t denotes the training set at time *t*. In our case, the dependent variable represents the number of nights spent by tourists in a given municipality according to register data at time *t*, while the features are chosen among different possible combinations of MNO data at time *t* and earlier and register data from some earlier time point.

3.2. Quasi-Transfer Learning

Transfer learning techniques are designed to improve predictive performance by incorporating knowledge from a related source model. In our case, we consider a Quasi-Transfer Learning (QTL) approach, which aims to enhance estimation by leveraging auxiliary models trained on similar but distinct populations.

Let $\mu(x, \beta)$ be a target model with unknown parameters β. Suppose we have access to a related source model, $\mu\left(x, \hat{\theta}\right)$, estimated separately on a different but similar population. Both models belong to the same family but differ in their parameter values, β and θ. Transfer learning, in this context, aims to enhance the estimation of β by leveraging $\hat{\theta}$. The term Quasi indicates that the method does not directly aim at estimating the target model but rather an approximation that is close to it (Zhang, 2024).

In our framework, we define q_t as:

$$q_t = s_t \cup r_t$$

where s_t represents the set of municipalities that have already provided the value of the response variable y_t, and r_t consists of the remaining units for which the response is not available yet. Given our specific application, we assume that all features (administrative and MNO) are available for r_t .

To account for the unavailability of some responses from the target units in the current period, we introduce an adjusted sample set s_t^*, defined as:

$$s_t^* = s_t \cup r_{t'}^*$$

where $r_{t'}^*$ represents a substitution mechanism in which the missing response values are replaced by the corresponding values from the previous year. In this step, some adjustments may be applied to the replacement values. In our case, to account for a trend effect primarily driven by COVID-19 recovery, we applied a +3% adjustment to the 2022 values.

Furthermore, we denote past observations at a previous time point *b* as s_b^*, q_b and we define the transfer schema:

$$\begin{tikzcd}
\mu(x, s_t^*) & & \mu(x, s_b^*) \arrow[ll, "\hat{g}(\cdot)"'] \\
& \Downarrow & \\
\mu(x, q_t^*) & & \mu(x, q_b^*) \arrow[ll, "\hat{g}(\cdot)"']
\end{tikzcd}$$

which applies the relationship g between s_t^* and s_b^* to q_t and q_b. In our work, we choose a transfer function $g(\cdot)$ that uses $\mu(x, s_b^*)$ as a feature, modelling the relation to be transferred as:

$$E\{\mu(x, s_t^*)\} = g\big(x, \mu(x, s_b^*)\big)$$

Where $\mu(x, s_t^*)$ represents the expected value of the response variable given x, and $\mu(x, s_b^*)$ corresponds to the source model. In summary, $\mu(x, s_b^*)$ and $\mu(x, q_b)$ are source models, and $\mu(x, q_t)$ is a target model; all of them share the same functional form but are trained on different data subsets, resulting in distinct parameter values. Alternative substitution mechanisms of unavailable response values, transfer schemes and functions can be considered depending on the specific application. For a detailed discussion, we refer to Zhang (2024).

3.3. *Augmented Learning*

Augmented learning (AL) techniques are designed to improve predictive performance by enhancing the training set, s, introducing some surrogate units. Given the notation already introduced for transfer learning, the training set is expanded with additional instances, r^*, that, in our case, correspond with the units in the target set r but the values of the dependent target variable $\{y_t \mid j \in r^*\}$ for these units are proxies for the target observations and are taken from some earlier time points. The main objective of this approach is to reduce the bias that arises from learning only from s, thereby improving the results of the baseline Simplistic Model previously described.

In this study, we choose to augment the sample s of fast-response municipalities by using, for the remaining units, the corresponding values from the same months of the previous year. The augmented sample s_t^* has already been described in the Quasi-Transfer Learning paragraph (Section 3.2). We also attempted to replicate the s_t units multiple times in the augmented sample to increase their weight in the training set, but we decided to augment it only once, as no significant improvements were observed.

4. Modeling

We applied AL, QTL, and Simplistic approach to predict the response variable for the months of September and October 2023. All the methodologies are implemented using both a Linear Regression Model (LM) and a Random Forest (RF). For each month we aimed to predict, and for each model, we selected specific variables to achieve the best results. The data available contain both spatial (cross sectional) and temporal information that could be exploited to make forecasting. For simplicity, we refer to the response variable, representing the nights spent as a tourist in a municipality, as *y*, and its MNO proxy as *x*. In the Linear Model, the square root transformation of both *y* and *x*, used as a feature, was applied to reduce the influence of extreme values and normalize the distribution of the variables. This approach improves the stability and accuracy of the model, especially considering that these variables exhibit a highly skewed distribution.

To find an adequate Linear Model, we started including the proxy *x* of the current month, and *x* auto-regressive up to lag 12. We then performed a backward elimination, keeping the most significant variables. Additionally, we included a categorical variable *Zone* as spatial information. The results of these procedures are the following models:

For September

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-3} + \beta_3 y_{t-12} + \beta_4 x_t + \beta_5 x_{t-1} + \beta_6 x_{t-9} + \sum_j \gamma_j \, Zone_j + \epsilon$$

For October

$$\begin{aligned} y_t = \beta_1 y_{t-1} + \beta_2 y_{t-3} + \beta_3 y_{t-6} + \beta_4 y_{t-8} + \beta_5 y_{t-12} + \beta_6 x_t + \beta_7 x_{t-1} + \beta_8 x_{t-2} \\ + \beta_9 x_{t-6} + \beta_{10} x_{t-7} + \beta_{11} x_{t-8} + \beta_{12} x_{t-9} + \sum_j \gamma_j \, Zone_j + \epsilon \end{aligned}$$

We applied the same procedure for Random Forest, performing backward selection based on feature importance. Since Random Forest is better suited than linear regression to handle a large number of features, we also included some register information. For both September and October, we selected the same set of features:

$$\left\{ \begin{matrix} y_{t-1}, y_{t-11}, y_{t-12}, x_t, x_{t-1}, x_{t-11,} x_{t-12}, Resident\_population, \\ Costal\_municipality, Altitude\_of\_center, Urbanization\_degree, \\ Extension\_km2, Province \end{matrix} \right\}$$

For our settings, we used the previous models for $\mu(x)$. For QTL, we obtained $g(x)$ by adding $\hat{\mu}(x)$ to the models and choosing b as the month preceding the one we wanted to predict. As stated in Section 3.3, for AL, we chose to augment the sample only once.

5. Preliminary Results

We tried to predict September and October 2023 overnight counts, pretending not to observe them from the administrative data, while observing them through MNO data (earlier available). We observed the admin data and MNO referring to previous months, and some characteristics of the municipalities. We applied AL, QTL and a Simplistic models to our data and compared their Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \ ,$$

along with the total predicted sum for the entire set of selected municipalities, which can be compared to the actual total. To evaluate the estimates for each available unit in the test set, we performed a 2-fold cross-validation. To reduce the influence of the sampling choice, we repeated this process 50 times. The reported MAPE values were computed as the average of all estimates obtained across these repeated cross-validations.

We decided to evaluate the algorithms on municipalities with a clear touristic vocation, selecting only those with a response value greater than the median, which in our case is 577 overnights. The results are shown in Table 2.

Table 2 – *Model results for September and October 2023.*

Model	MAE (Sep)	MAPE (Sep)	TOT ŷ (Sep)	MAE (Oct)	MAPE (Oct)	TOT ŷ (Oct)
Linear (Simplistic)	1301.94	19.93	4025369	1486.04	20.94	1766437
Linear (QTL)	1206.17	19.04	4006008	1205.74	20.47	1805782
Linear (AL)	1227.11	19.26	4009297	1238.69	20.35	1807020
Random Forest (Simplistic)	4707.84	27.12	4144826	4625.61	35.31	2389966
Random Forest (QTL)	2356.2	19.73	4104355	1900.56	21.94	1945381
Random Forest (AL)	2224.88	18.98	4092401	1860.68	21.52	1941391
TOT y			4079632			1798189

We also computed 95% Confidence Intervals (CI) and normalized Confidence Intervals, defined for a municipality *i* as:

$$CI_{normalized,i} = \frac{CI_{upper,i} - CI_{lower,i}}{\hat{\bar{y}}_i}$$

where $\hat{\bar{y}}_i$ denotes the mean of the predictions for municipality *i* across all repetitions. By taking the mean of the $CI_{normalized}$ for each unit across the QTL, AL, and Simplistic models, we obtained the results given in Table 3.

Table 3 – *Mean CI Normalized for Simplistic, QTL, and AL Models.*

	LINEAR MODEL			RANDOM FOREST		
	Simplistic	QTL	AL	Simplistic	QTL	AL
September 2023	0.222	0.104	0.104	0.346	0.114	0.084
October 2023	0.301	0.097	0.098	0.384	0.087	0.055

In Figure 2, for the month of September, the left panel shows the ratio between the absolute error of the Simplistic model and that of the other two models (Linear Model and Random Forest) for each municipality. The right panel displays the corresponding ratios for the widths of the Confidence Intervals.

In Figure 3, for the month of September, we report the MAPE of the Linear Model for the municipalities where the ratio between the absolute error of the Simplistic model and that of the more advanced methods (AL and QTL) exceeds 1.5. Figure 4 presents the same analysis for the Random Forest model. The aim of these figures is to highlight the municipalities with the lowest MAPE among those where the advanced methods outperform the Simplistic model.

Figure 2 – *September 2023 QTL vs Simplistic.*

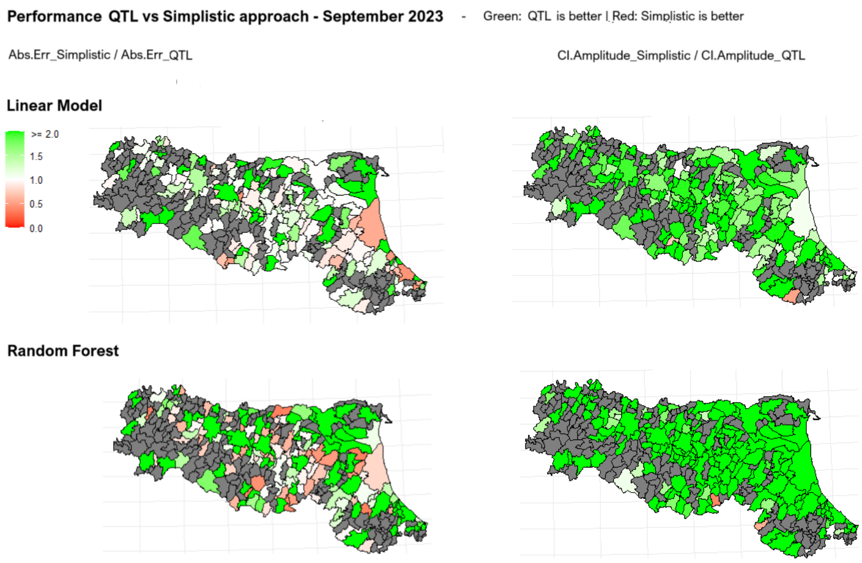

Performance QTL vs Simplistic approach - September 2023 - Green: QTL is better | Red: Simplistic is better
Abs.Err_Simplistic / Abs.Err_QTL
CI.Amplitude_Simplistic / CI.Amplitude_QTL
Linear Model
>= 2.0
1.5
1.0
0.5
0.0
Random Forest

Figure 3 – *QTL and AL MAPE - Linear Model.*

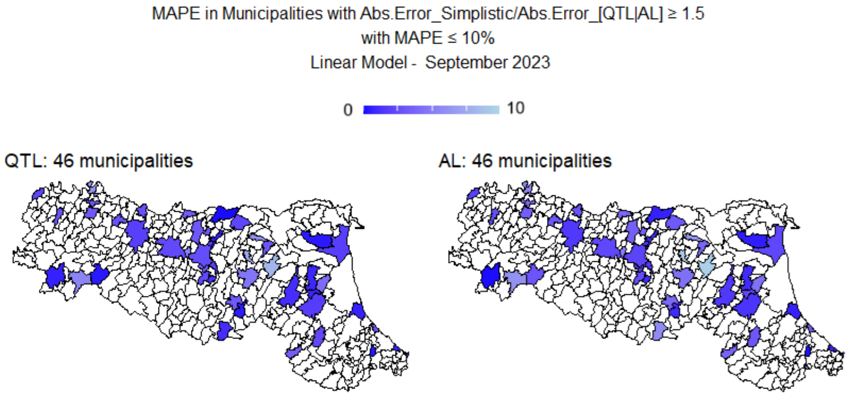
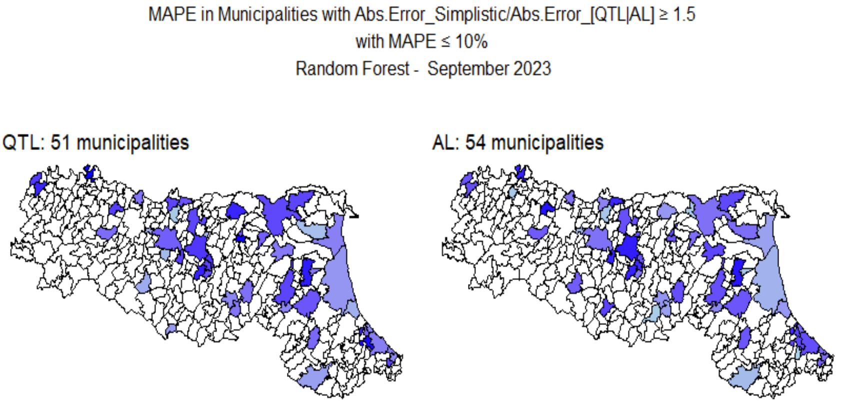

MAPE in Municipalities with Abs.Error_Simplistic/Abs.Error_[QTL|AL] ≥ 1.5
with MAPE ≤ 10%
Linear Model - September 2023
0
10
QTL: 46 municipalities
AL: 46 municipalities

Figure 4 – *QTL and AL MAPE – Random Forest.*


MAPE in Municipalities with Abs.Error_Simplistic/Abs.Error_[QTL|AL] ≥ 1.5
with MAPE ≤ 10%
Random Forest - September 2023
QTL: 51 municipalities
AL: 54 municipalities

Observing the results, we can state that QTL and AL improve the precision of the estimates. Linear models appear to be more suitable for this prediction task, but Random Forest with QTL and AL achieves significant improvements over the Simplistic approach, reaching or even surpassing the MAPE of its counterpart, which benefits less from these advanced methods. Additionally, as shown in Table 3 and in Figures 3-4 Confidence Interval plots, AL and QTL effectively improve estimates' stability.

6. Models Ensemble

We observed that the Random Forest model performed better in municipalities with low tourism levels (fewer than 5,000 overnight stays per month), while the linear model was more effective for highly sparse data with large, regression-dominating values. Then, we implemented a simple ensemble of the two models: for municipalities where the observed target variable in the previous year exceeded a predefined threshold, we used linear regression for predictions, while Random Forest was employed for all cases below this threshold.

The fixed threshold is 1300; in this application the threshold was chosen empirically to minimise the MAPE. Approximately two-thirds of municipalities were predicted using the linear model, while the remaining third were predicted with Random Forest. The ensemble results are shown in Table 4 and 5.

Table 4 – *Ensemble Model results for September 2023.*

	September 2023		
Model ensemble	MAE	MAPE	$TOT_{\hat{y}}$
Simplistic	1315.20	22.78	4043084
QTL	1198.93	19.19	4014355
AL	1214.54	18.76	4017478
$\boldsymbol{TOT_y}$			4079632

Table 5 – *Ensemble Model results for October 2023.*

	October 2023		
Model ensemble	MAE	MAPE	$TOT_{\hat{y}}$
Simplistic	1500.68	23.23	1778993
QTL	1194.13	19.15	1810638
AL	1228.30	19.35	1813292
$\boldsymbol{TOT_y}$			1798189

It is important to note that the global percentage error we obtained (approximately 19%), is not uniformly distributed across the territory. Specifically, for about 40% of municipalities' predictions deviate from the true values by less than 10%.

We need to further investigate municipalities where MNO (Mobile Network Operator) data and administrative data show poor correlation, particularly those with significantly overestimated MNO values compared with administrative records. For instance, Bentivoglio municipality exhibits percentage errors of 143% and 189% in September and October, respectively. This discrepancy may arise because Bentivoglio hosts a major hospital with a high volume of patient admissions, which the MNO data could misinterpret as tourist presence.

7. Concluding Remarks and Future Improvements

Our study highlights several aspects that need to be investigated and developed to improve forecasts.

In our experiments, we observed that the linear model performed better in predicting overnight stays in municipalities with a high number of them, whereas the random forest yielded better results in municipalities with fewer overnight stays. Based on this observation, an ensemble of the two models has been considered. Additionally, a mixed strategy could be explored, selecting among the simplistic, Quasi Transfer Learning, and Augmented Learning approaches, depending on which method performs best in a given municipality. To implement this, appropriate ensemble criteria and selection methods need to be investigated.

In this work, we tested Quasi Transfer Learning and Augmented Learning with the linear model and random forest, but many other models could be suitable for our forecasting task, such as count models (e.g., negative binomial), neural networks, and ARIMA. Further experiments could also be conducted using additional register data, which may enhance the identification of long-term dependencies while adjusting for short-term variations using MNO data, which are only available for recent periods.

In this study, we adopted the implementation choices that we considered most appropriate for our context, but Zhang and Haug (2024) propose several alternative mechanisms for substituting unavailable response variables, as well as transfer schemes and functions that could be further explored.

References

ZHANG L.-C., HAUG J. K. 2024. Turnover flash estimation by purposive sampling and debit card transactions, *Journal of Official Statistics*, SAGE Publications.

Cristina FARICELLI, Istat, cristina.faricelli@istat.it
Alessandro PIOVANI, Istat, alessandro.piovani@istat.it
Tiziana TUOTO, Istat, tuoto@istat.it