Tourist Mobility Forecasting with Region-based Flows and Regular Trips

Fernando Terroso-Saenz
Technical University of Cartagena
Cartagena, Spain
e-mail: fernando.terroso@upct.es

Juan Morales-García
Universidad Católica de Murcia (UCAM)
Murcia, Spain
e-mail: jmorales8@ucam.edu

Miguel Puig
Universidad Católica de Murcia (UCAM)
Murcia, Spain
e-mail: mpuig@ucam.edu

Ginesa Martinez-del Vas
Universidad Católica de Murcia (UCAM)
Murcia, Spain
e-mail: gmvas@ucam.edu

Andres Muñoz
University of Cadiz
Cadiz, Spain
e-mail: andres.munoz@uca.es

***Abstract*—One of the most prominent courses of action in the tourist sector is the development of predictors to anticipate the flow of incoming and outgoing tourists of a region. To do so, most of the existing approaches usually take tourist-related flows as the only primary input to perform the prediction. The present work assesses the suitability of composing a deep-learning predictor that fuses touristic displacements with data extracted from a general-purpose human-mobility dataset. The proposal has been tested in the Region of Murcia, a Spanish administrative area with a lively tourist sector. Results show that our approach achieves up to 46% Root Mean Square Error (RMSE) reduction with respect to a baseline only relying on tourist data.**

Keywords-tourist mobility ; deep neural networks , human mobility flows ; time series forecasting

I. Introduction

The tourism sector has undergone extensive research aimed at devising intelligent solutions to enhance both business processes and customer experiences in multiple platforms, such as online social networks [1]. This integration has facilitated the analysis of tourist mobility behavior. Consequently, forecasting tourists' flows holds significant implications in areas such as tourism marketing or services, empowering tourism institutions and stakeholders to more efficiently manage their resources [2][3].

However, the advancement of prediction algorithms to forecast tourist flows (e.g., the volume of incoming or outgoing tourist trips within a geographical area such as a city) typically depends on a *univariate* approach, where the target flow serves as the primary input for the predictor [4][5]. Yet, the utilization of alternative forms of human mobility data as *exogenous* variables to enhance prediction accuracy has not been thoroughly investigated.

The primary objective of this study is to evaluate the viability of enhancing a tourist-flow prediction model with human movement data obtained from sources that document regular and daily movements within a specific geographical area. This concept is rooted in the notion that daily human movements towards a region could offer an alternative yet supplementary perspective on its tourist flows. For instance, a significant increase in inbound tourists to a city attributed to a social event might be preceded by a decrease in commuter journeys towards that region several hours before the event begins. Anticipating such a decline could be leveraged by a predictive model to enhance the accuracy of forecasting future tourist visitation patterns.

In order to evaluate our approach, we have used several instances of a model comprising a stack of convolutional and recurrent neural network layers for time series forecasting in order to anticipate different types of touristic flows towards the Region of Murcia (RM), a Spanish Administrative area in Spain with an active tourist industry. In that sense, the predictor is feed with different subflows extracted from an open nationwide human-mobility dataset to anticipate the overall number of incoming tourists towards RM several weeks ahead.

The salient contribution of this work is the fusion of different datasets that allows the development of a touristic mobility predictor that merges the regular and tourist movement of a geographical region so as to forecast its incoming touristic flow. The key benefit of this *multi-flow* approach is that it allows a much more accurate estimation of the tourists arriving to a region than an approach solely relying on tourist-related flows for several time horizons.

The remainder of the paper is structured as follows. Section 2 summarizes the most relevant current approaches for human mobility prediction in the touristic sector. Then, Section 3 describes the use-case setting of our study. In Section 4, the most important results of the deployed palette of predictors are described and evaluated. Lastly, Section 5 summarizes the main conclusions and potential future research lines motivated by this work.

II. Related Work

In recent years, there has been a large interest in harnessing methodologies to fuse data from heterogeneous sources so as to forecast tourist movements, thus enhancing tourism planning and management through the utilization of multivariate datasets. One approach has been based on the usage of tourist flows from one region to improve the prediction accuracy in another area. For instance, Zhu et al. [6] examined tourist flows from six countries to forecast tourist arrivals in Singapore, proposing a pairwise modeling approach to account for interdependence among countries within the same geographic region. Analyzing

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-212-8

flows from 1995 to 2013 and predicting up to 20 quarters ahead, they demonstrated improvements in Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) by incorporating pairwise flows compared to models treating flows independently, particularly evident in annual predictions. A similar approach was followed by Yang et al. [7] investigated the combination of spatial and temporal tourist flow datasets in China, contrasting univariate models like ARIMA with multivariate space-time autoregressive moving average (STARMA) models. Their study, spanning tourist mobility data from 29 Chinese regions between 1987 and 2016, showcased enhanced accuracy of STARMA models, especially in neighboring regions with strong spatial correlations.

Another line of research has focused on the integration of datasources not directly related to human mobility to enhance the prediction performance. As a matter of fact, Zhang et al. [8] integrated tourism flow volumes with various Search Intensity Indicators (SII) from Google Trends to refine the accuracy of Machine Learning and Deep Learning models in forecasting tourist arrivals in Hong Kong over different time horizons. A similar approach was adopted for predicting tourist demand in Macau, China, by Law et al. [9], who evaluated diverse Deep Learning models, particularly those employing attention mechanisms, surpassing conventional ML techniques like Support Vector Regression. Besides, De-Jesus et al. [10] made use of data reporting the evolution of the COVID-19 pandemic in the Philippines to enhance the prediction of inbound tourist to that country confirming that integrating such type of data improved the model accuracy.

The novel aspect of our current work lies in the nature of data used to enhance tourist mobility prediction. Unlike previous approaches relying on web-based indicators, or COVID-19 data as exogenous variables, we leverage direct human movement data extracted from an open and general-purpose feed specific to the geographical area of interest.

III. Use-Case Setting

The focal point of our investigation has been the Region of Murcia (RM), an autonomous community in Spain situated in the southeast of the country (see Figure 1). This area boasts a population of approximately 1.5 million people and covers an expanse of 11,313 km^2. In terms of tourism, this region welcomed over 1,300,000 visitors in 2022, marking a 45% increase compared to 2021 [11].

A. Datasets

We utilized two distinct mobility datasets, the first encompasses tourist mobility within the Region of Murcia (RM), while the second encompasses total human mobility within the same region. This way, we had two different views of how people moved around the target region.

1) Tourist Mobility Dataset (TMD): The flow of tourists in RM was gathered through the Tourist Mobility Dataset (TMD) provided by the Tourism Institute of the Region of Murcia [12] as part of its Smart Region project. This dataset captures the inbound and outbound movement of tourists in RM over a 16-month period spanning from January 1st, 2022, to April 30th, 2023. These flows are derived from the network events generated by mobile phones connected to the Telefonica network, one of the leading carriers in Spain [13]. The data undergoes anonymization, extrapolation, and aggregation stages to compute the final tourist flows included in the dataset.

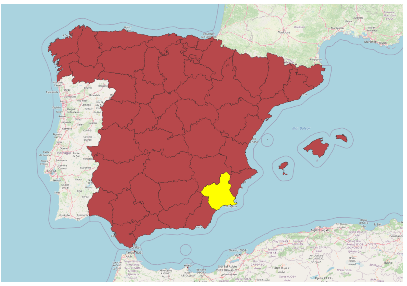


Figure 1. Location of the Region of Murcia (in yellow) with respect to the rest of autonomous communities in Spain depicted in red.

An important aspect of this dataset is its distinction between the incoming flow of national (residing in Spain) and international (arriving from other countries) tourists ($\mathcal{NT}$ and $\mathcal{IT}$, respectively), as well as excursionists ($\mathcal{NE}$ and $\mathcal{IE}$, respectively). The former category comprises individuals who spend at least one night in the region (e.g., staying in a hotel, camping, or tourist accommodation), while the latter encompasses day trippers who visit RM for the day but do not spend the night away from their primary residence.

The format of the dataset defines the flows on a weekly basis. In this manner, the mobility of the w-th week of the year y for a city c is defined as a single tuple,

$$\langle y, w, c, f^m_{work}, f^a_{work}, f^n_{work}, f^m_{end}, f^a_{end}, f^n_{end} \rangle$$

where f^m_{work}, f^a_{work} and f^n_{work} are the *week slices* comprising the overall incoming flows towards c during the morning (m), afternoon (a) and night (n), respectively, considering all the working days of the w-th week. Similarly, f^m_{end}, f^a_{end} and f^n_{end} provide the same time-sliced flows for the weekend days (Saturday and Sunday in Spain). Thus, for each combination of year, week and city (y, w, c) the dataset comprises 4 different tuples, one for each type of touristic flow, 1) national excursionists $\mathcal{NE}$, 2) national tourists $\mathcal{NT}$, 3) international excursionists $\mathcal{IE}$ and 4) international tourists $\mathcal{IT}$. For the sake of clarity, Figure 2 shows the number of incoming tourists and excursionists (regardless its origin) for the 70 weeks covered by the dataset.

Given the aforementioned flows, we computed 3 aggregated ones comprising the overall number of tourists $\mathcal{T}$ $(= \mathcal{NT}+\mathcal{IT})$, the overall number of excursionists $\mathcal{E}$ $(= \mathcal{NE} + \mathcal{IE})$ and the overall number of visitors $\mathcal{A}$ (=$\mathcal{T} + \mathcal{E}$).

Next, we composed a timeseries for each flow $\mathcal{F} \in \langle \mathcal{NE}, \mathcal{NT}, \mathcal{IE}, \mathcal{IT}, \mathcal{T}, \mathcal{E}, \mathcal{A} \rangle$ covering the 70 weeks under study with the following format $\mathcal{F}_{TM} = \langle f^{m,1}_{work} \rightarrow f^{a,1}_{work} \rightarrow f^{n,1}_{work} \rightarrow f^{m,1}_{end} \rightarrow f^{a,1}_{end} \rightarrow f^{n,1}_{end} \rightarrow f^{m,2}_{work} \rightarrow f^{a,2}_{work} \rightarrow f^{n,2}_{work} \rightarrow f^{m,2}_{end} \rightarrow f^{a,2}_{end} \rightarrow f^{n,2}_{end} \rightarrow ... \rightarrow f^{m,70}_{work} \rightarrow f^{a,70}_{work} \rightarrow$

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-212-8

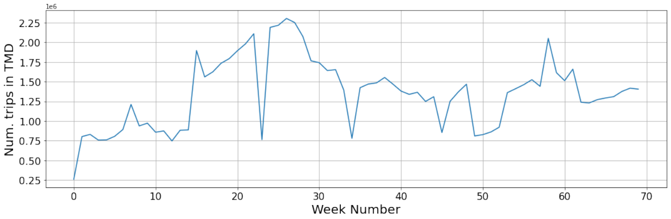


Figure 2. General flow of incoming tourists and excursionists to the Region of Murcia during the period of study considering the tourist mobility dataset.

$f^{n,70}_{work} \rightarrow f^{m,70}_{end} \rightarrow f^{a,70}_{end} \rightarrow f^{n,70}_{end} \rangle$ where, for example, $f^{m,i}_{work}$ is the record comprising the overall value of the $\mathcal{F}$ flow during the working days' mornings of the i-th week.

2) General Human Mobility Dataset (GMD): This dataset was obtained from the nationwide human mobility report published by the Spanish Ministry of Transportation (SMT) in January 2022 [14]. It provides information on the number of trips per hour between 2,735 cities across Spain, covering both the mainland and insular regions. This dataset can be viewed as a collection of tuples, each taking the form,

$$\langle date, hour, m_{origin}, m_{dest}, n_{trp}, dist \rangle$$

reporting that there was n_{trp} human trips from the city m_{origin} to the city m_{dest} and whose distance was $dist$ km during the indicated $date$ and $hour$.

As per official reports [15], these mobility data were derived from Call Detail Records (CDRs) of 13 million users from an undisclosed mobile carrier. After anonymization, the dataset was utilized to extrapolate comprehensive mobility patterns representative of the Spanish population at a national scale, subsequently released as open data. It is important to note that this dataset encompasses the movements of individuals irrespective of their mode of transportation.

Utilizing this dataset, we filtered its flows by retaining records that satisfy the following two criteria: (1) their destination m_{dest} is one of the cities in RM, and (2) their distance $dist$ exceeds a certain threshold δ. The first criterion captures the inbound flows to RM, while the second criterion refines these inbound flows based on the distance traveled by visitors to reach RM. Since the threshold δ defines the minimum distance, we avoid the fact that some of the resulting flows include almost all the records of the initial dataset.

Considering the peninsular area of Spain has an approximate radius of 540km and commuters' average travel distances range from 19 to 34 km [16], we employ three distinct δ values: 100, 400, and 800 km. Using these thresholds, we initially derived a subset of short-distance trips with $\delta = 100km$ ($\mathcal{F}^{100}_{GM}$), aiming to capture regular and non-touristic trips alongside various types of tourist flows to RM. Subsequently, we constructed a second subset representing medium-distance trips with $\delta = 400km$ ($\mathcal{F}^{400}_{GM}$) and a third subset encompassing long-distance travelers with $\delta = 800km$ ($\mathcal{F}^{800}_{GM}$). This approach allowed us to progressively filter out the proportion of regular and non-touristic trips in each subset by increasing the value of δ. In that sense, each subflow is defined as a timestamped sequence $\mathcal{F}^{\delta}_{GM} = \langle f^{m,1}_{gm,\delta,work} \rightarrow f^{a,1}_{gm,\delta,work} \rightarrow f^{n,1}_{gm,\delta,work} \rightarrow f^{m,1}_{gm,\delta,end} \rightarrow f^{a,1}_{gm,\delta,end} \rightarrow f^{n,1}_{gm,\delta,end} \rightarrow f^{m,2}_{gm,\delta,work} \rightarrow f^{a,2}_{gm,\delta,work} \rightarrow f^{n,2}_{gm,\delta,work} \rightarrow f^{m,2}_{gm,\delta,end} \rightarrow f^{a,2}_{gm,\delta,end} \rightarrow f^{n,2}_{gm,\delta,end} \rightarrow ... \rightarrow f^{m,70}_{work} \rightarrow f^{a,70}_{gm,\delta,work} \rightarrow f^{n,70}_{gm,\delta,work} \rightarrow f^{m,70}_{gm,\delta,end} \rightarrow f^{a,70}_{gm,\delta,end} \rightarrow f^{n,70}_{gm,\delta,end} \rangle$ where, for example, $f^{m,i}_{gm,\delta,work}$ is the record comprising the overall value of the trips towards RM, covering a distance of at least δ km, during the morning of the working days of the i-th week according to the GMD.

Figure 3 shows the time series of the aforementioned subflows in RM during the same time period considered for the touristic dataset (2022/01/01-2023/04/30). As observed, the 3 time series have a very different order of magnitude. Furthermore, $\mathcal{F}^{100}_{GM}$ exhibits a quite *flat* pattern throughout the whole period of study capturing a mostly-stationary travel behaviour. This is consistent with the fact this GMD flow is the one that captures more regular and commuting trips from the three ones. On the contrary, flows $\mathcal{F}^{400}_{GM}$ and $\mathcal{F}^{800}_{GM}$ comprised sharper peaks during the different holiday seasons covered in the study. For example, Figure 3 shows a large increment of incoming trips in both flows during the two Easter holidays under consideration. This reveals that these two timeseries captured more *seasonal* travel patterns and, thus, more compatible with tourist-related displacements.

As we can see, each GMD subflow provides different point of view of the human mobility in the Region of Murcia so that they can be used to study whether the usage of general human mobility might improve the prediction of a particular flow of visitors. It is important to remark that the GMD and SMT represent different mobility flows. However, some redundancy might ocurre between both datasets as the GMD might comprise a certain number of touristic trips.

IV. Description of the Predictor

The focus of this paper lies on addressing the tourist mobility prediction challenge, which can be framed as a regression problem:

Given the weekly time slice w, the number of incoming tourists and/or excursionists over the past w_{prev} time slices according to the TMD $\mathcal{F}^{w}_{TM} = \langle f^{w}, f^{w-1}, .., f^{w-w_{prev}, \mathcal{S}} \rangle$, and the number of incoming trips based on the GMD within a distance-threshold δ for the same time lags $\mathcal{F}^{\delta,w}_{GM} = \langle f^{w}_{gm,\delta}, f^{w-1}_{gm,\delta}, .., f^{w-w_{prev}}_{gm,\delta} \rangle$, **Determine** a mapping function $\mathcal{P}$:

$$\mathcal{P}(\mathcal{F}^{w}_{TM}, \mathcal{F}^{\delta,w}_{GM}) \rightarrow \mathcal{F}^{w+T}_{TM}$$

where $\mathcal{F}^{w+T}_{TM}$ represents the estimated number of tourists and/or excursionists arriving in RM at the $(w+T)$th week slice as per the TMD study, with T denoting the prediction time horizon ($T \geq 1$). Notably, the novelty in this predictive model lies in its integration of GMD-based trips, supplementing the information derived solely from the TMD source.

To implement the predictor model, we have used a combination of a Convolutional Neural Network (CNN) with a Long short-term model (LSTM), resulting in a CNNLSTM model. As Figure 4 shows, this model firstly compresses

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

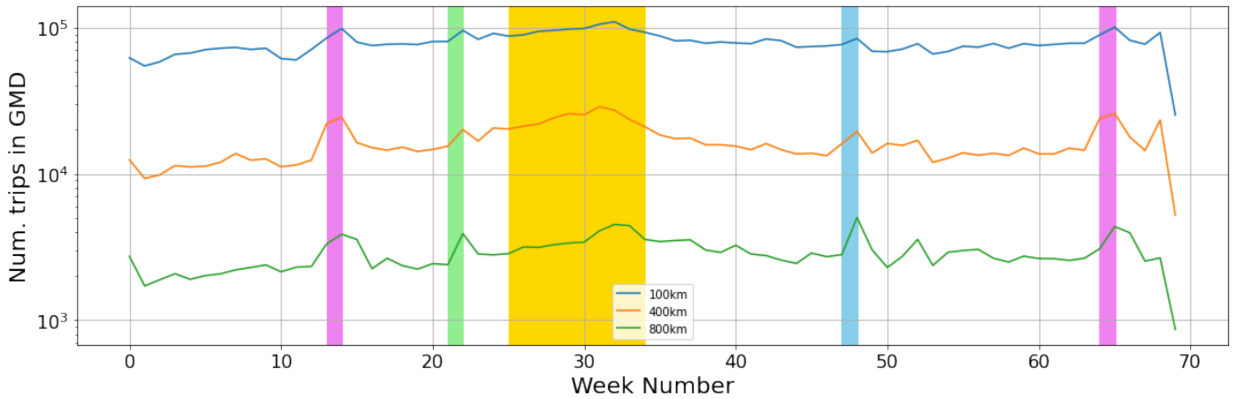


Figure 3. General flow of incoming tourists and excursionists to the Region of Murcia, considering different distance-based filtering (δ) during the period of study considering the general human mobility dataset. The violet areas represent the Easter holidays in 2022 and 2023 respectively, the green and blue ones nation bank holidays and the yellow area the summer period (July and August) in 2022.

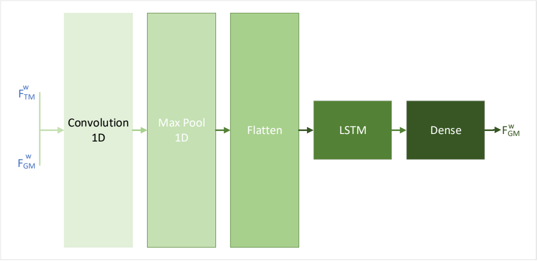


Figure 4. Layer architecture of the CNNLSTM applied in the study.

and extracts the relevant features of the incoming bi-variate timeseries comprising $\mathcal{F}^{w}_{TM}$ $\mathcal{F}^{\delta,w}_{GM}$ flows by means of a one-dimensional convolutional and a max-pool layer. Then, the resulting sequence is *flattened* to a 1D vector in order to be processed by the downstream LSTM and dense layers and generate the estimated $\mathcal{F}^{w+T}_{TM}$ flow. In that sense, we opted to use a CNN instead of applying feature selection prior to an LSTM because CNNs are capable of efficiently capturing spatial and local patterns in time series data, which helps identify complex relationships between the data before being processed by the LSTM. This approach enables better extraction of relevant features directly from the time series, eliminating the need for prior manual feature selection.

V. Evaluation of the Predictor

In this section we evaluate the accuracy of the CNNLSTM predictor described in the previous section.

A. Metrics

In terms of evaluating the CNNLSTM model, the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) [17] stand out as two widely utilized metrics for assessing accuracy in predicting continuous variables. These metrics are well-suited for comparing models as they quantify the average prediction error of the model in the units of the variable of interest. Their definitions are as follows,

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i,)^2},$$

where, for our experiment, y_i is the real number of touristic trips, $\hat{y}_i$ is the predicted number of trips and n is the number of observations. Furthermore, we complement these metrics with the Mean Average Prediction Error (MAPE) metric that is calculated as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}| \times 100.$$

B. Model Hyper-parameters

Table I comprises the configuration of the hyper-parameters applied in the CNNLSTM model.

C. Evaluation Results

Table II shows the metric values of the CNNLSTM model for different combination of inputs. In order to properly evaluate the impact of our approach, the table also shows the results of a baseline model that is only fed with touristic mobility data ($\mathcal{F}_{TM}$). Bearing in mind the description of the predictor stated in Section IV, this baseline can be defined as a univariate function $\mathcal{P}(\mathcal{F}^{w}_{TM}) \rightarrow \mathcal{F}^{w+T}_{TM}$.

As observed in Table II, at least one of the models enriched with general mobility data outperformed the univariate alternative for almost all the target flows. For example, in order to predict the overall flow of visitors to RM ($\mathcal{A}$), the predictor fed with the GMD flow with a 800km distance threshold ($\mathcal{A}$, $\mathcal{F}^{800}_{GM}$) reduced the MAE from 30,678.195 to 29,311.016 (-5%). A higher improvement is observed in the case of the national excursionists ($\mathcal{NE}$), in which the RMSE dropped from 15,732.939 in the univariate version to 8,544.438 (-46%) in the model using the $\mathcal{F}^{800}_{GM}$ flow. Concerning the $\mathcal{F}^{400}_{GM}$ flow, it

TABLE I
HYPER-PARAMETERS OF THE CNNLSTM MODEL

Hyperparameter	Description	Value
Filters	Features detector	64
Kernel size	Filters matrix used to extract the features from the dataset	2
Strides	Number of timesteps shifts over the input sequence	4
Activation function	Function that decide if a neuron should be (or not) activated	Tanh
Batch size	Size of batch used for training/forecasting	32
Epochs (+ *EarlyStopping*)	Number of epochs used in training	15000
Optimizer	Function that optimises the learning of a artificial intelligence model	Adam
Loss function	Function used for evaluate the error of the model in each epoch	MSE
Learning rate (+ *ReduceLROnPlateau*)	Percentage change with which weights are updated at each iteration	0.003
Train-test split	Rate of the dataset used to train and evaluate the models	90% (train), 10% (test)

allowed to improve the accuracy of the prediction of the $\mathcal{NE}$ flow (MAPE from 35.145 to 16.443, -54%) and the $\mathcal{T}$ flow (MAE from 16,909.334 to 15,519.060, -9%).

An important finding of this evaluation is that the most suitable distance threshold δ to compose the GMD flow varies depending on the target tourist flow. This makes sense as the nature of each target flow is quite different. For example, a 400km-δ provided a higher accuracy for the national flows, $\mathcal{NE}$ and $\mathcal{NT}$, whereas the 800km provided the best results for the international-tourist flow ($\mathcal{IT}$). Moreover, the 400km threshold was the best configuration to predict the overall touristic flow ($\mathcal{T}$) but the 800km one allowed the best accuracy for the overall excursionist ($\mathcal{E}$). This dichotomy of distances for tourists and excursionists explain why the most accurate model to predict the global flow $\mathcal{A}$ depends on the evaluation metric under consideration as we can see in the first 4 rows of Table II.

It is important to remark that the predictor actually improved its accuracy when it incorporated the regular flows filtered with δ equals to 400 or 800km. However, the $\mathcal{F}_{GM}^{100}$ did not provide a clear improvement for any of target flows. In that sense, $\mathcal{F}_{GM}^{400}$ and $\mathcal{F}_{GM}^{800}$ were the two flows exhibiting a higher seasonality with peaks at certain holiday periods revealing that the *weight* of the touristic displacements was quite high in such flows (Section III-A2). This suggests that the actual improvement of the predictor occurs when it is enriched with an exogenous flow comprising *latent* touristic displacements in a quite strong manner. That is, when it provides an *alternative* view of the touristic flows of RM discarding the regular displacements at high degree.

Furthermore, Figure 5 shows the RMSE of each model for each target flow and different values of time horizons T, namely 6, 12, 18, 24, 32 and 48 week slices. Since the TMD defines 6 slices per week (Section III-A1), such time horizons can be also regarded as 1, 2, 3, 4, 5.3 and 8 weeks. As observed, the major improvement of the CNNLSTM with GMD flows usually occurred for large time horizons above 24 slices (4 weeks). This is specially noticeable in the $\mathcal{A}$ (Figure 5a), $\mathcal{E}$ (Figure 5b) and $\mathcal{NE}$ (Figure 5d) flows. It is also worth mentioning the fact that the model enriched with $\mathcal{F}_{GM,800}$ clearly outperformed the other models for all the time horizons in order to predict the $\mathcal{IT}$ flow. This is consistent with the fact that predictors enriched with GMD data could learn the latent mobility patterns from several points of view and, thus, anticipate their long-term behaviour in a more accurate manner.

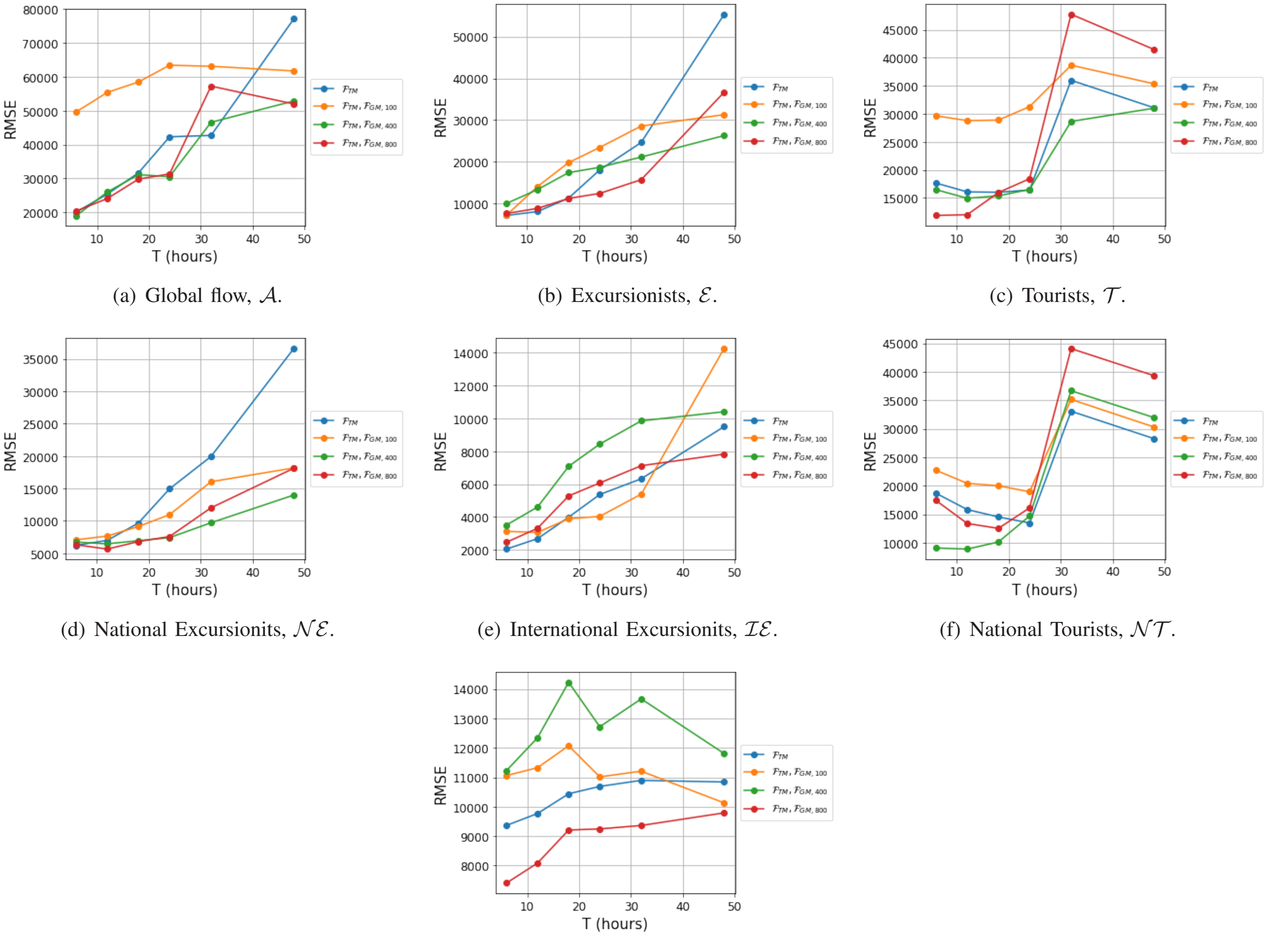
TABLE II
ERROR METRICS OF EVALUATED MODELS. THE VALUES IN BOLD INDICATE THE LOWEST ERROR FOR EACH ⟨TARGET FLOW, METRIC⟩ PAIR.

Target flow	Model's input	MAE	RMSE	MAPE
$\mathcal{A}$	$\mathcal{A}$	30678.195	39895.783	13.764
$\mathcal{A}$	$\mathcal{A}$, $\mathcal{F}_{GM,100}$	44274.756	58651.799	20.762
$\mathcal{A}$	$\mathcal{A}$, $\mathcal{F}_{GM,400}$	29402.056	**34364.585**	12.206
$\mathcal{A}$	$\mathcal{A}$,$\mathcal{F}_{GM,800}$	**29311.016**	35820.294	**11.998**
$\mathcal{IE}$	$\mathcal{IE}$	**3816.539**	**4983.614**	24.589
$\mathcal{IE}$	$\mathcal{IE}$, $\mathcal{F}_{GM,100}$	3939.599	5628.872	**20.193**
$\mathcal{IE}$	$\mathcal{IE}$, $\mathcal{F}_{GM,400}$	5980.346	7317.516	37.581
$\mathcal{IE}$	$\mathcal{IE}$, $\mathcal{F}_{GM,800}$	4370.739	5344.350	26.865
$\mathcal{NE}$	$\mathcal{NE}$	12739.819	15732.936	35.154
$\mathcal{NE}$	$\mathcal{NE}$, $\mathcal{F}_{GM,100}$	9769.504	11509.648	25.019
$\mathcal{NE}$	$\mathcal{NE}$, $\mathcal{F}_{GM,400}$	**6871.259**	**8544.438**	**16.443**
$\mathcal{NE}$	$\mathcal{NE}$, $\mathcal{F}_{GM,800}$	7461.158	9421.957	16.706
$\mathcal{E}$	$\mathcal{E}$	15197.112	20773.676	29.911
$\mathcal{E}$	$\mathcal{E}$, $\mathcal{F}_{GM,100}$	15496.672	20741.639	33.164
$\mathcal{E}$	$\mathcal{E}$, $\mathcal{F}_{GM,400}$	14815.044	17828.300	28.326
$\mathcal{E}$	$\mathcal{E}$, $\mathcal{F}_{GM,800}$	**11095.700**	**15434.727**	**20.736**
$\mathcal{IT}$	$\mathcal{IT}$	8403.148	10333.727	12.709
$\mathcal{IT}$	$\mathcal{IT}$, $\mathcal{F}_{GM,100}$	10044.294	11134.676	15.101
$\mathcal{IT}$	$\mathcal{IT}$, $\mathcal{F}_{GM,400}$	10709.579	12664.216	15.613
$\mathcal{IT}$	$\mathcal{IT}$, $\mathcal{F}_{GM,800}$	**7694.914**	**8850.736**	**11.922**
$\mathcal{NT}$	$\mathcal{NT}$	16799.757	20668.949	14.330
$\mathcal{NT}$	$\mathcal{NT}$, $\mathcal{F}_{GM,100}$	21503.516	24611.753	18.729
$\mathcal{NT}$	$\mathcal{NT}$, $\mathcal{F}_{GM,400}$	**13661.015**	**18604.120**	**10.718**
$\mathcal{NT}$	$\mathcal{NT}$, $\mathcal{F}_{GM,800}$	18360.056	23826.051	14.729
$\mathcal{T}$	$\mathcal{T}$	16909.334	22205.017	9.350
$\mathcal{T}$	$\mathcal{T}$, $\mathcal{F}_{GM,100}$	27295.051	32113.587	15.807
$\mathcal{T}$	$\mathcal{T}$, $\mathcal{F}_{GM,400}$	**15519.060**	**20499.031**	**8.806**
$\mathcal{T}$	$\mathcal{T}$, $\mathcal{F}_{GM,800}$	17648.112	24587.760	9.052

Finally, the aforementioned evaluation shows that the usage of alternative human trips actually improved the prediction of most of the touristic flows under consideration. However, this improvement actually occurred when the GMD flow, $\mathcal{F}_{GM}^{\delta}$, comprised seasonal mobility that was compatible with the touristic activity rather than reporting regular and commuting trips. Besides, this improvement was most noticeable as long as the target time horizon increased.

VI. CONCLUSIONS AND FUTURE WORK

The utilization of human mobility data is revolutionizing the tourism industry, enabling the development of predictive models to optimize resource allocation for hotel companies.



(a) Global flow, $\mathcal{A}$.

(b) Excursionists, $\mathcal{E}$.

(c) Tourists, $\mathcal{T}$.

(d) National Excursionits, $\mathcal{NE}$.

(e) International Excursionits, $\mathcal{IE}$.

(f) National Tourists, $\mathcal{NT}$.

(g) International Tourists, $\mathcal{IT}$.

Figure 5. RMSE per time horizon for each target flow.

However, many existing models overlook general mobility patterns beyond tourist flows. In this study, we propose integrating general mobility data to enhance the accuracy of tourism flow predictions. Our innovative approach combines Convolutional Neural Network and Long-Short Term Memory models, enabling us to forecast tourist flows up to 8 weeks ahead with increased precision.

Testing our methodology on data collected from the Region of Murcia (Spain) over a 16-month period demonstrates significant improvements in accuracy, with error reductions exceeding 50%. This underscores the potential of integrating general mobility data into existing predictive models to better anticipate tourist behaviors.

Two avenues of research can be explored further in subsequent stages of this study. Firstly, the integration of additional contextual data, such as weather conditions, events, and holidays could further enhance the predictive accuracy of the tourist flow model. Secondly, expanding the analysis to encompass multiple regions would offer a broader understanding of tourist flows at a national level. This might reveal intricate patterns and dynamics between different areas, thereby enriching the comprehensiveness of tourist mobility forecasting.

Acknowledgements

Financial support for this research has been provided by the Instituto de Turismo de la Región de Murcia through project CAT/TU/47-22 "Cátedra Internacional de Inteligencia Turística Región de Murcia UCAM-ITREM".

References

[1] B. Armutcu, A. Tan, M. Amponsah, S. Parida, and H. Ramkissoon, “Tourist behaviour: The role of digital marketing and social media”, *Acta psychologica*, vol. 240, p. 104 025, 2023.

[2] G. S. Atsalakis, I. G. Atsalaki, and C. Zopounidis, “Forecasting the success of a new tourism service by a neuro-fuzzy technique”, *European Journal of Operational Research*, vol. 268, no. 2, pp. 716–727, 2018.

[3] H. Albuquerque, C. Costa, and F. Martins, “The use of geographical information systems for tourism marketing purposes in aveiro region (portugal)”, *Tourism management perspectives*, vol. 26, pp. 172–178, 2018.

[4] C. Li, P. Ge, Z. Liu, and W. Zheng, “Forecasting tourist arrivals using denoising and potential factors”, *Annals of Tourism Research*, vol. 83, p. 102 943, 2020. DOI: 10.1016/j.annals.2020.102943.

[5] W. Wang *et al.*, “A multi-graph convolutional network framework for tourist flow prediction”, *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 4, pp. 1–13, 2021. DOI: https://doi.org/10.1145/3424220.

[6] L. Zhu, C. Lim, W. Xie, and Y. Wu, “Modelling tourist flow association for tourism demand forecasting”, *Current Issues in Tourism*, vol. 21, no. 8, pp. 902–916, 2018.

[7] Y. Yang and H. Zhang, “Spatial-temporal forecasting of tourism demand”, *Annals of Tourism Research*, vol. 75, pp. 106–119, 2019.

[8] Y. Zhang, G. Li, B. Muskat, and R. Law, “Tourism demand forecasting: A decomposed deep learning approach”, *Journal of Travel Research*, vol. 60, no. 5, pp. 981–997, 2021.

[9] R. Law, G. Li, D. K. C. Fong, and X. Han, “Tourism demand forecasting: A deep learning approach”, *Annals of tourism research*, vol. 75, pp. 410–423, 2019.

[10] N. M. De Jesus and B. R. Samonte, “AI in tourism: Leveraging machine learning in predicting tourist arrivals in philippines using artificial neural network”, *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, pp. 1–8, 2023.

[11] Government of the Region of Murcia, *Tourism Statistics - Region of Murcia*, https://www.turismoregiondemurcia.es/es/estadisticas_de_turismo/, Accessed: October 8, 2024, 2024.

[12] Government of the Region of Murcia, *Smart Tourism Destination (DTI) - ITREM*, https://www.itrem.es/nexo/dti/, Accessed: October 8, 2024, 2024.

[13] Telefónica S.A., *Telefónica - Telecommunications Company*, https://www.telefonica.com/es/, Accessed: October 8, 2024, 2024.

[14] Ministry of Transport, Mobility and Urban Agenda, *Daily basic mobility study using Big Data*, https://www.mitma.es/ministerio/proyectos-singulares/estudios-de-movilidad-con-big-data/estudio-basico-diario, Accessed: October 8, 2024, 2024.

[15] M. Secretariat of State for Transport and U. Agenda, “Analysis of Mobility in Spain Using Big Data Technology During the State of Alarm for Managing the COVID-19 Crisis”, Spainsh Ministry of Transport, Mobility and Urban Agenda, Tech. Rep., 2020.

[16] G. Pasaoglu *et al.*, “Travel patterns and the potential use of electric cars–results from a direct survey in six european countries”, *Technological Forecasting and Social Change*, vol. 87, pp. 51–59, 2014.

[17] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”, *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005, cited By (since 1996)149.