Forecasting Supply in Voronoi Regions for App-Based Taxi Hailing Services

Ravina Gelda
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai 600 036, India
Email: ee14s045@ee.iitm.ac.in

Krishna Jagannathan
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai 600 036, India
Email: krishnaj@ee.iitm.ac.in

Gaurav Raina
Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai 600 036, India
Email: gaurav@ee.iitm.ac.in

***Abstract*—In this paper, we deal with the problem of supply forecasting in the context of an application based taxi hailing service. We first propose a method to optimally partition the city space using a Voronoi tessellation. The generating points of the Voronoi regions are obtained as demand density cluster centers, from the taxi demand dataset. We also identify the optimal temporal resolution to use for forecasting supply in these Voronoi regions. We use a linear time-series based algorithm to forecast supply in each Voronoi region. Using this methodology for the city of Bengaluru, India, we obtained a supply forecast accuracy of about 90% for the heavily used Voronoi regions. This represents a substantial improvement in the forecast accuracy compared to similar time-series based approaches, employed over rectangular 'geohashes'.**

***Index Terms*—Transportation, GPS data, Spatio-temporal data mining, Time series analysis, Voronoi tessellation, Forecasting.**

I. Introduction

In most big cities, taxis constitute an important mode of transport. A major challenge faced by taxi services is the imbalance between supply and demand. The problem is aggravated by the fact that demand is inherently ad hoc, and taxi drivers (supply) are mostly self-interested and cannot be controlled centrally. Thus, while attaining an appropriate supply-demand equilibrium is important for profitability as well as reliability, achieving a supply-demand balance is quiet challenging. In particular, from the viewpoint of drivers, knowledge about the undersupplied areas can be advantageous, especially when it is not economically viable to adopt random cruising strategies.

Since demand hotspots are of great interest for the company and the drivers, partitioning the city into demand density clusters, and predicting the supply in these zones is useful in tackling the supply-demand mismatch problem. Further, we should be able to predict supply for a particular area before the status of the area (under-supplied / over-supplied) changes. In this paper, we consider the problem of supply forecasting, for a mobile application based taxi hailing service. We propose a forecasting algorithm that uses a time-series based approach at an optimally chosen spatio-temporal resolution. Specifically, our method involves identifying the demand hotspot centers, and generating a Voronoi tessellation of the city, with the demand density cluster centers as the generating points of the Voronoi regions. We then use a suitable linear time-series model for each Voronoi region, to forecast future supply. Our approach shows an improved forecast accuracy, when compared to similar time-series based techniques [1] employed over 6-level geohahses (i.e., a 1.2km$\times$0.6km rectangular area). We believe the improved accuracy obtained in this paper is due to the optimal partitioning of the city into demand density clusters.

A. Related Work

Multiple works in the literature have already explored the taxi GPS data with distinct applications like modeling the spatio-temporal structure of taxi services [13], predicting the taxi-passenger demand [8] or building intelligent passenger-finding strategies [5]. The work presented in [11] is closely related to ours – therein, the authors propose a model to predict the number of vacant taxis in a $1 \times 1 km^2$ based on time of the day, day of the week, and the weather condition. In [7], the authors propose a forecasting model using ARIMA to predict the number of services that will emerge at a given taxi stand. In [8], the authors propose a learning model to predict the real-time spatio-temporal distribution of the taxi-passenger demand using perceptrons. Despite the useful insights these papers provide, they do not address the problem of choosing optimal spatio-temporal resolution, *i.e.*, the effect of varying spatio-temporal resolutions on the prediction accuracy. In this work, we focus on predicting the short-term supply by using time-series models, with an effective spatio-temporal resolution.

In a recent paper [1], we proposed a forecasting methodology using a linear time-series framework to predict the supply for the next 24 hours. We used a temporal resolution of 1 hour and spatial resolution of 6-level geohash, and forecast the supply with an accuracy of approximately 76%. Further, by exploiting the spatial correlation between neighbouring 6-level geohashes, we could improve the forecast accuracy to about 89% for the most heavily used 6-level geohashes.

In the above context, a couple of important questions arise: (i) Is there an optimal way of spatially partitioning the city in order to attain a higher forecast accuracy and spatio-temporal resolution? (ii) What is the optimal spatio-temporal resolution for employing the time-series models?

In this paper, we propose a technique of partitioning the city into Voronoi regions, with demand density centers as

the generating points. It has been noted in the literature that Voronoi tessellation allows for a fairly realistic representation of the dynamic spatial fields [9]. We use GPS traces of taxi trips for 2 months from the city of Bengaluru, India for analysis and evaluation. A popular clustering algorithm, the K-means clustering, is used for finding the demand clusters and their centers. We choose 800 cluster centres for K-means clustering of the demand, so that the city gets partitioned into regions of approximate area $0.9km^2$ on an average. Next, in order to choose the optimal temporal resolution, we compare the accuracy of one day ahead (24 hours) predictions with a baseline model of the seasonal näive method [3, Chapter 2, Section 2.3] for different time scales. We choose, a well-known error metric, SMAPE (Symmetric Mean Absolute Percentage Error), for comparing the forecast accuracy. However, since this metric can be too stringent when errors are made in predicting actual values that are small in magnitude, we use a modified version of the SMAPE – by adding a Laplace estimator to the original definition [4]. For the data set used in this paper, the optimal temporal resolution turns out to be around 30 to 60 minutes for most of the Voronoi regions. We compare our approach using Voronoi tessellation with the 6-level geohash approach in [1] for supply forecasting. This comparison is made at different time scales. Our results show that the Voronoi tessellation method for supply forecasting results in more accurate forecast than the 6-level geohash method, for all the time scales considered. In particular, for 30 to 60 minutes time scale, the improvement in accuracy of supply forecast with the baseline model considered, ranges from 12% to 33%. Next, we predict the supply (number of vacant taxis) for the Voronoi regions of the city of Bengaluru, India, using the forecasting methodology described in [1]. We are able to predict the supply with 90% accuracy for heavily used Voronoi regions for one day ahead predictions.

In summary, our contributions are as follows:

1) We propose a technique of partitioning the city space using Voronoi regions, with demand density cluster centers as the generating points of the tessellation.
2) We identify the optimal temporal resolution for forecasting taxi supply, corresponding to the Voronoi tessellation proposed above.
3) We show, using empirical results, that the Voronoi tessellation method significantly outperforms the 6-level geohash method of spatial partitioning, for the purpose of accurately forecasting supply.

The rest of the paper is organized as follows. Section II consist of an introduction to the data used in this paper along with the preprocessing techniques used. In Section III, we describe the procedure for the Voronoi tessellation along with the empirical results for the data used in this paper, followed by the method for optimal temporal resolution selection along with the empirical results in Section IV. Therein, we also compare the Voronoi tessellation method with the 6-level geohash method for spatial partitioning of the city for supply forecasting. In Section V, we forecast the supply using appropriate linear time-series forecasting models and present the forecast results. Finally, we conclude the paper in Section VI.

II. Data Preprocessing

In this section, we describe the dataset used in this paper and also the data preprocessing techniques used. We worked with a rich dataset collected from GPS devices of nearly 14 million taxi trips, spanning over a period of two months (January and February). The data was acquired from a leading Indian transportation company dealing with app-based taxi rental services. It contains the taxi booking demand generated by the public, along with user identification numbers, location in the form of latitude and longitude pair and the time of booking. It also contains the driver identification number, along with the driver status, location in the form of latitude and longitude pair and the time of login.

If a customer ID is detected multiple times in 30 minutes in $1km^2$ area, it is taken as a single booking demand because it is unlikely that a customer will make multiple bookings in 30 minutes, from the same $1km^2$ area. Similarly, if a driver ID is detected multiple times in 30 minutes in $1km^2$ area, it is taken as a single login session because it is unlikely that a driver will become available within 30 minutes from the same $1km^2$ area.

We used 'R- A statistical tool' [14] for data analysis, simulation and modelling. In order to perform K-means clustering as well as Voronoi tessellation, Euclidean distance was used as the similarity measure on the taxi GPS data used in this paper. To that end, we translate the latitude and the longitude to UTM coordinate system. UTM (Universal Transverse Mercator) is an international location reference system that depicts the Earth's three-dimensional surface in a relatively accurate, two-dimensional manner.

III. Optimal spatial resolution selection

In this section, we describe the technique of spatial division of the city into Voronoi regions.

Voronoi tessellation is a method of partitioning a plane into convex polyhedra such that each polyhedra contains exactly one generating point, and every point in a given polyhedra is closer to its generating point than to any other generating point. The boundary of two regions is the set of points which are equidistant from the shared generating points [10]. In order to partition the city space, we have to find appropriate generating points. These generating points are chosen as the locations of interest to the transportation company as well as the drivers – typically, locations of demand density cluster centers.

A. Demand Clustering

We use K-means clustering for finding demand density clusters. The K-means algorithms is a well known and simple unsupervised learning algorithm that is widely used to determine hotspots [2, Chapter 13, Section 13.2.1]. The algorithm offers a simple method to classify a given dataset into a certain pre-specified number (say k) of clusters. The main idea is to define k centers, one for each cluster. These centers

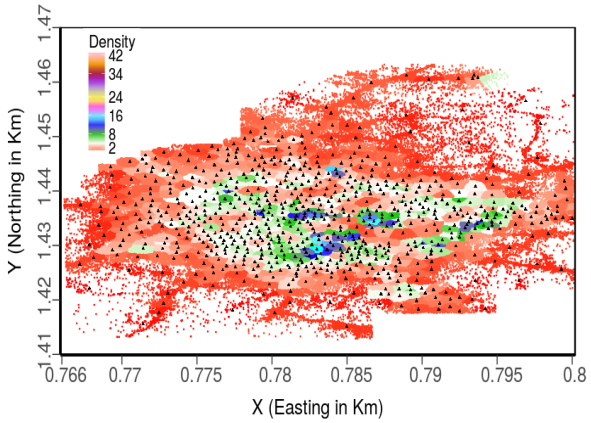


Fig. 1: The heatmap of demand density in 800 areas of the city of Bengaluru, India with black dots representing the demand density cluster center.

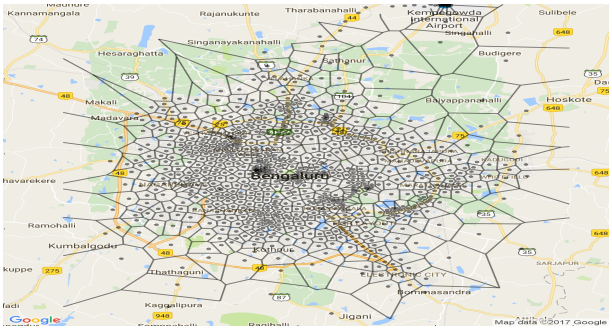


Fig. 2: Voronoi tessellation superimposed over the city map of Bengaluru. Black dots represent the tessellation centers.

should be initialised in a judicious manner because changing their locations can cause widely different results. Therefore, it would be a good choice to place them as far apart as possible [12].

The following are the key steps in the K-means clustering algorithm:

1. Select k initial centers for clustering.
2. Associate each point in the given dataset to its nearest center to form clusters.
3. Recalculate k new centers as the barycenters [2, Chapter 13, Section 13.2.1] of the clusters formed in the previous step.
4. Form new clusters by associating the data points to nearest cluster centers calculated in the previos step.
5. If there is a reassociation of any data point in step [4], then go to step [3]. Else, stop.

We use the K-means clustering algorithm on the taxi demand data from Bengaluru to obtain the demand density cluster centers, which will later serve as the generating points for the Voronoi tessellation. Let $\{z_i\}_{i=1}^{n}$ be the locations of the demands in the city of Bengaluru, obtained after data preprocessing (see Section II). Here, $\{z_i = (x_i, y_i), i \in \{1,2,\ldots,n\}\}$, and (x_i, y_i) are the UTM coordinates of the i^{th} location of the demand. Bengaluru city has an area of approximately $741Km^2$, which we aim to divide into 800 Voronoi cells. This ensures that the city gets partitioned into regions of roughly $0.9Km^2$ each, on an average. Therefore, k is chosen to be 800 for K-means clustering.

The objective function to minimize for the K-means algorithm is

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||z_i^{(j)} - c_j||^2. \quad (1)$$

The initial cluster centers, denoted by $\{c_j\}_{j=1}^{800}$, are chosen as the centers of the 6-level geohashes belonging to the popular areas in the city of Bengaluru. Note that $z_i^{(j)}$ represents the demand location nearest to the cluster center c_j. In the objective function in Equation (1), $||z_i^{(j)} - c_j||^2$ represents the squared distance between a data point $z_i^{(j)}$ and the cluster center c_j. Denote by $\left\{c_j^*\right\}_{j=1}^{800}$ the final demand cluster centers when the K-means algorithm converges.

The demand density of a cluster is considered to be proportional to the number of demands in that cluster and inversely proportional to the distance between cluster center and the farthest point in that cluster. To be more precise, let $\mathbb{L} = \{l_j\}_{j=1}^{800}$ be the number of demands in clusters $\left\{c_j^*\right\}_{j=1}^{800}$. Denote by $\mathbb{M} = \{m_j\}_{j=1}^{800}$ the distance of the farthest point in the cluster from the center, i.e., $m_j = \max_{1 \le a \le l_j} ||z_a^{(j)} - c_j^*||^2$. The density of demand clusters, $\mathbb{D} = \{d_j\}_{j=1}^{800}$, is then defined as

$$d_j = \frac{l_j}{m_j}.$$

Fig. 1 shows the heatmap of the resulting demand clusters, along with the cluster centers. These clusters are shaded according to the demand density in that cluster.

B. Voronoi tessellation with demand density cluster centers

We use cluster centers obtained from K-means clustering for Voronoi tessellation. These demand cluster centers are used as the generating points for the Voronoi tessellation of the city of Bengaluru, based on the Euclidean distance. The Voronoi region V_j corresponding to the point c_j^* is defined by:

$$V_j = \left\{ x \in \mathbb{R}^2 | \quad ||x - c_j^*|| \leqslant ||x - c_i^*|| \quad \forall j \ni i \neq j \right\}, \quad (2)$$

where $||.||$ is the Euclidean distance.

The resulting Voronoi tessellation of the city space of Bengaluru is shown in Fig. 2. Next, we define 'top Voronoi' regions as the Voronoi regions where most of the demand (64% of the total demand) is concentrated. These top Voronoi regions comprise a total of 255 Voronoi regions. The average area of the top Voronoi regions is found to be $0.494Km^2$. Fig. 3 shows a histogram of the area of the top 255 Voronoi regions.

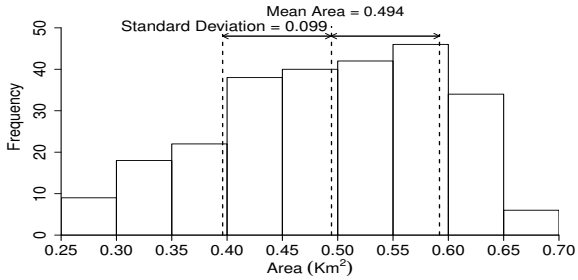


Fig. 3: Distribution of the area of the top Voronoi regions.

IV. Optimal temporal resolution selection

In this section, we describe the method for optimal temporal resolution selection for supply forecasting. We also compare the performance of the Voronoi tessellation with the performance of the 6-level geohash method in [1].

In order to find the optimal time scale as the aggregation period for time-series analysis and supply prediction, we consider factors like accuracy of prediction for test data, accuracy of model fitting for training data, usefulness of the prediction in real-time actions, irregularity (noise) in time-series, as well as computational cost.

We formulate a time-series for supply data of 2 months for each top Voronoi region. The forecasting model is trained using the first 59 days of the time-series and tested for the 60^{th} day. The baseline model of seasonal näive method is considered for comparing the forecast accuracy of supply time-series of Voronoi regions for different aggregation periods. A modified version of the frequently used percentage error metric, Symmetric Mean Percentage Error (SMAPE) is used to test the forecast accuracy of the time-series. This metric is discussed below.

Denote by p the aggregation periods in minutes considered in this paper; in particular $p \in \{5,15,30,60,120,300,720\}$. Let $\{y^v_{t,p}\}_{v=1}^{255}$ denote the aggregate supply for all the top 255 Voronoi regions for an aggregation period p in the interval $[t,t+p]$. A time-series with an aggregation period p for 2 months (60 days) period for a v^{th} Voronoi region can be formulated as $y^v_p = \left\{ y^v_{0,p}, y^v_{1,p}, \ldots y^v_{t,p}, \ldots y^v_{\lceil \frac{86400}{p} \rceil, p} \right\}$.

Baseline model: The Seasonal näive method [3, Chapter 2, Section 2.3] is considered as the baseline model. In this model, we perform averaging over all the past seasons. That is,

$$y_{t+1,p} = \frac{1}{N} \sum_{i=1}^{N} y_{t+1-im,p}, \tag{3}$$

where N is the total number of seasons, and m is the seasonality period.

Modified SMAPE: SMAPE is a well-known error metric employed to evaluate forecast accuracy [6]. It is defined as follows:

$$SMAPE = \frac{100}{N} \sum_{i=1}^{N} \frac{|F_i - A_i|}{F_i + A_i}. \tag{4}$$

However, it has the disadvantage of being infinite or undefined if there are zero values in a time-series, as is frequent for intermittent data. To produce more accurate statistics about time-series containing very small numbers, we may add a Laplace estimator [4] to the original definition of SMAPE in Equation (4). This is accomplished by adding a constant c to the denominator, as used in the similar context in [8]. With this, modified SMAPE can be defined as follows:

$$SMAPE = \frac{100}{N} \sum_{i=1}^{N} \frac{|F_i - A_i|}{F_i + A_i + c}. \tag{5}$$

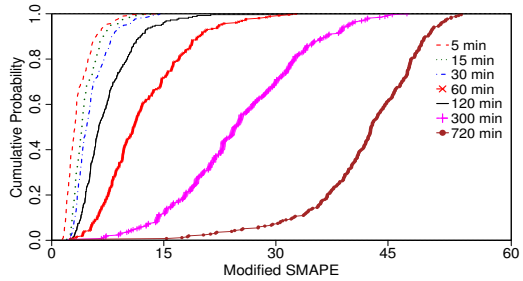
Here, A_i is the i^{th} actual value belonging to a time-series and F_i is the forecast value for i_{th} position of a time-series. In our case $c = 1$, the most used value [4]. N is the number of predicted values of a time-series.

Observations: In order to compare the accuracy of forecast for all the aggregation periods considered, we compare the Empirical Cumulative Distribution Function (ECDF) of the modified SMAPE. We calculate modified SMAPE for one day ahead prediction of supply with the baseline model for each aggregation period for the top Voronoi regions. From Fig. 4a, we observe that the ECDF curves reach unity faster as the aggregation periods decreases. Hence, we may conclude that the supply for the top Voronoi regions is predicted with higher accuracy as the aggregation period decreases. Next, we plot ECDFs of modified SMAPE calculated for baseline model fit for the train data of the top Voronoi regions for each aggregation period considered (See Fig. 4b). From Fig. 4b, we observe that the ECDF curves reach unity faster as the aggregation periods decreases. Hence, we may conclude that the baseline model fit for train data of the top Voronoi regions improves as the aggregation period decreases.

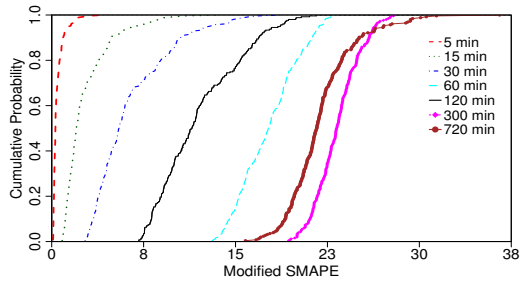
In order to compare the irregularities (noise) present in the supply time-series for different aggregation periods, we plot normalised supply data for one day for Bengaluru for aggregation periods considered, as shown in Fig. 5. We observe that irregularities are present in time-series with lower aggregation periods of 5 and 15 minutes. In addition to that the shape of the observations is not preserved for time-series with 120, 300, and 720 minutes aggregation period. Hence, we may conclude that it is difficult to capture the features of data for time-series with 120, 300, and 720 minutes aggregation period.

As the aggregation period of the time-series increases, one would have to train the time-series model for more samples. This leads to an increase in the computational cost. For example, for an aggregation period of 5 minute, 59 days of training data contributes to $16992 = (12 \times 59 \times 24)$ samples and for 60 minutes as the aggregation period, 59 days of training data contribute to $1416 = (1 \times 59 \times 24)$ samples.

Forecasting supply for the top Voronoi regions with the baseline model for lower aggregation periods of 5 and 15 minutes, is at most 4% more accurate than 30 to 60 minutes aggregation periods (see Fig. 6a). On the other hand, these smaller aggregation periods introduce irregularities in the time-series, in addition to incurring high computational cost.



(a) Empirical CDF of the modified SMAPE calculated for one day ahead predictions at different aggregation periods for the top Voronoi regions.



(b) Empirical CDF of the modified SMAPE calculated for baseline model fit for 59 days train data of the top Voronoi regions at different aggregation periods.

Fig. 4: Empirical CDF of the modified SMAPE calculated for test and train data for the top Voronoi regions at different aggregation periods.

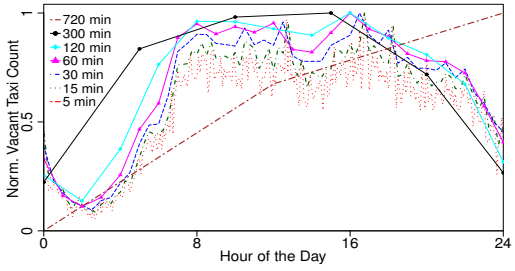
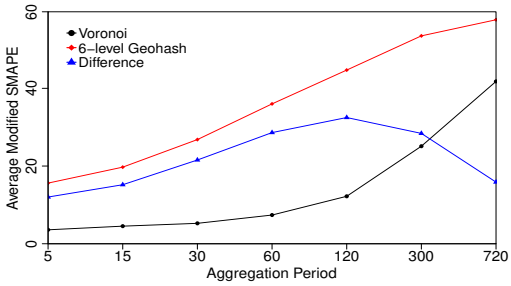


Fig. 5: Normalized supply time-series aggregated over different time-scales.

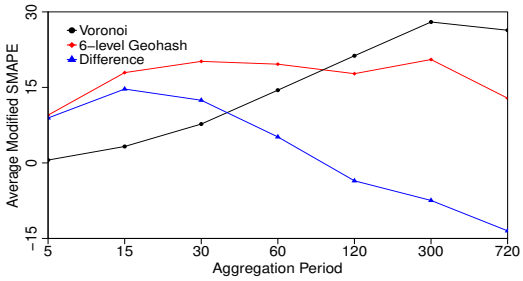
Forecasting supply for the top Voronoi regions using 120, 300, or 720 minutes aggregation periods is not useful for real-time actions. In addition, the time-series features such as seasonality could not be captured at these coarse time scales, and the forecast accuracy decreases (see Fig. 6a). Taking all these factors into account, we choose 30 to 60 minutes aggregation periods as being well suited for supply forecasting. At these time scales, the forecast accuracy is fairly good, and the features of the time-series could be captured well with low computational cost.

A. Comparison of Voronoi tessellation versus 6-level geohashes using the Baseline model

Similar to the definition of top Voronoi regions in Section III-B, we define top 6-level geohashes as the 6-level geohashes in which 64% of demand is concentrated. In our



(a) Comparision of average modified SMAPE for one day ahead prediction of supply.



(b) Comparision of average modified SMAPE for baseline model fitting with supply train data.

Fig. 6: Comparision of average modified SMAPE calculated for test data and train data for different aggregation periods for top Voronoi regions and top 6-level geohashes.

dataset, this corresponds to 80 top 6-level geohashes for Bengaluru.

We formulate a time-series for supply data of 2 months for each top 6-level geohash. The forecasting model is trained using the first 59 days of the time-series and tested for the 60^{th} day. Let $\left\{y^g_{t,p}\right\}_{g=1}^{80}$ denote the aggregate supply for all the top 80 6-level geohashes for an aggregation period p in the interval $[t, t+p]$. A time-series with an aggregation period p for 2 months (60 days) period for a g^{th} 6-level geohash can be formulated as $y^g_p = \left\{y^g_{0,p}, y^g_{1,p}, \ldots y^g_{t,p}, \ldots y^g_{\lceil \frac{86400}{p} \rceil, p}\right\}$.

Observations: The spatial resolution for 6-level geohashes is $0.72Km^2$, where as that for the Voronoi regions, $0.464Km^2$, on an average (see Fig. 3).

Now, we calculate the modified SMAPE for one day ahead prediction of supply with the baseline model (3) for each top 6-level geohash, for all the aggregation periods considered. From Fig. 6a, we observe that the difference between average modified SMAPE for the top 6-level geohashes and the top Voronoi regions ranges from $11\% - 33\%$ for all the aggregation periods considered. Also, we observe that for each aggregation period, the average modified SMAPE for the top Voronoi regions is lesser than that of the top 6-level geohashes. Next, we compare the modified SMAPE for the accuracy of fit for the training data, under the baseline model, for the top 6-level geohashes (see Fig. 6b). We observe that the difference between average modified SMAPE for the top 6-level geohashes and the top Voronoi regions is in the range -13% – 15% for all the

aggregation periods considered. Even though the training error for the top 6-level geohashes is lower than that for the top Voronoi regions for some aggregation periods, the one day forecast error is higher for 6-level geohashes. This is perhaps because of the more pronounced overfit/underfit issues in the 6-level geohash case in the training data.

From the above results, we conclude that the Voronoi partitioning leads to a more accurate forecast at a higher spatial resolution, for the same time resolution.

V. Comparision of Time-Series Models for Supply Forecasting

In this section, we briefly outline the comparision between various linear time-series models for supply forecasting over Voronoi regions. A more detailed description of the time-series models considered can be found in [1].

We analyse the time-series for the top Voronoi regions for an aggregation time period of 60 minutes to find the most suitable time-series model for prediction. Analysis of the time-series reveals the presence of strong seasonal periodicities in the supply data. There have been numerous proposals for forecasting models in the literature, such as TBATS, DSHW and STL decomposition [3, Chapter 6] which are suited to model this type of data. We also consider simple forecasting models like the average method and the seasonal näive method to ensure that the selected model is better than these simple alternatives [3, Chapter 2, Section 2.3]. Specifically, we compare among TBATS, DSHW, STL, ARIMA, SARIMA and the above simple alternatives, and choose a model which gives the least modified SMAPE.

The shortlisted models are those that resulted in a low modified SMAPE. In Fig. 7, we plot the ECDF of modified SMAPE calculated for different linear time-series models considered for forecasting supply in the top Voronoi regions. Results obtained show that the STL model [1] and the baseline (3) model perform better than the other time-series models considered. Also, the STL model performs better than the baseline model for some top Voronoi regions and vice-versa. After forecasting supply for the top Voronoi regions with the model that performs best for each top Voronoi region, we plot the ECDF of the resulting modified SMAPEs in Fig. 7. We were able to predict supply for all the top Voronoi regions with almost 90% accuracy.

VI. Concluding Remarks

In this paper, we presented an application of time-series models to forecast supply for the city of Bengaluru, India. We proposed a method for optimal spatial partitioning of the city using Voronoi tessellation. The generating points for Voronoi tessellation were obtained as the centers of the demand density clusters. We used K-means clustering algorithm to obtain demand density cluster centers from the demand dataset. We also identified 30 to 60 minutes time period as the optimal temporal resolution for forecasting supply in the top Voronoi regions obtained. We showed that supply forecast with the baseline model considered is more accurate for the top Voronoi cells than the top 6-level geohashes for all the aggregation periods considered. Our approach shows substantial improvement in accuracy of about 14% as compared to the similar time-series based approaches, employed over 6-level geohashes [1].

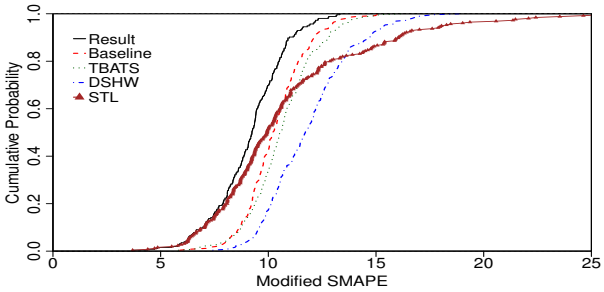


Fig. 7: Empirical CDF of the modified SMAPE calculated for the different forecasting models considered along with the final results obtained for the chosen models.

References

[1] R. Gelda, K. Jagannathan and G. Raina, "Taxi dispatches using supply forecasting: A time-series based approach", in *Proceedings of The* 14^{th} *IEEE International Conference on SmartCity*, 2016.

[2] R. Hastie, R. Tibshirani and J .Friedman, "The elements of statistical learning", *Berlin: Springer series in statistics*, 2001.

[3] R.J. Hyndman and G. Athanasopoulos, "Forecasting: principles and practice", 2016. [Online]. Available: https://www.otexts.org/book/fpp.

[4] E.T. Jaynes, Probability theory: The logic of science, *Cambridge University press*, 2003.

[5] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi and Q. Yang, "Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset", in *Proceedings of The IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2011.

[6] S. Makridakis and M. Hibon, "The M3-Competition: results, conclusions and implications", *International journal of forecasting*, vol. 16, pp. 451-476, 2000.

[7] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira and L. Damas, "Predicting taxi-passenger demand using streaming data based", in *Proceedings of The* 15^{th} *International IEEE Conference on Intelligent Transportation Systems*, 2015.

[8] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira and L. Damas, "On predicting the taxi-passenger demand: A real-time approach", in *Proceedings of The Portuguese Conference on Artificial Intelligence,* 2013.

[9] M.A. Mostafavi, L.H. Beni and K. Hins-Mallet, "Representing dynamic spatial processes using voronoi diagrams: Recent developments", in *Proceedings of The Sixth International Symposium on Voronoi Diagrams*, 2009.

[10] A. Okabe, B. Boots, K. Sugihara, and S.N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd ed.* Chichester, U.K, Wiley, 2000.

[11] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti, "Taxi-aware map: Identifying and predicting vacant taxis in the city", *Ambient Intelligence*, vol. 6439, pp. 86-95, 2010.

[12] J. Xu and H. Liu, "Web user clustering analysis based on KMeans algorithm", in *Proceedings of International Conference on Information, Networking and Automation (ICINA)*, 2010.

[13] Y. Yue, Y. Zhuang, Q. Li and Q. Mao, "Mining time-dependent attractive areas and movement patterns from taxi trajectory data", in *Proceedings of The* 17^{th} *IEEE International Conference on Geoinformatics*, 2009.

[14] R: A language and environment for statistical computing, 2016. [Online]. Available: http://www.R-project.org/.