Cars and Calls: Using CDR Data to Approximate Official Traffic Counts

Tony Liang
Computer Science Department
University of Maryland, College Park
tony@umd.edu

Vanessa Frias-Martinez
College of Information Studies
University of Maryland, College Park
vfrias@umd.edu

ABSTRACT

Official traffic counts approximate the amount of traffic observed in roads. These counts are computed by local authorities to model traffic and understand transportation needs. Although there exist automatic traffic collection techniques such as CCTVs or road sensors, these tend to be highly expensive. Thus, countries with limited resources typically compute the official traffic counts with manual approaches including individuals counting the number of cars that go through different roads. Given the ubiquity of cell phones in Senegal, we propose the use of cell phone data as a proxy for modeling traffic. Specifically, we design a technique to automatically compute official traffic counts using mobility features extracted from Call Detail Records. We evaluate the technique against official traffic count numbers in Senegal and obtain correlation coefficients between real and predicted values of $r = 0.889$. Our approach provides a reliable technique to measure traffic counts at large-scale and in affordable manner.

1. INTRODUCTION

Traffic counts approximate the amount of traffic observed on a particular road. These counts are critical for transportation planning and transportation analysts since many policy decisions, including the construction of new roads, are based on such numbers. Currently, there exist two types of data collection techniques to gather information regarding traffic counts: manual and automatic. Manual approaches involve hiring a number of observers who manually count the number of vehicles that drive through specific roads. This approach is not only expensive, but also very hard to scale to cover the bulk of highways and roads in a country. On the other hand, there exists a handful of automatic techniques for monitoring traffic including road sensors, cameras, or toll information among others. These techniques automatically gather road and traffic information with high frequency and generally report average values through specific periods of time e.g., daily averages. Although automatic techniques are easier to scale than manual approaches, they tend to be expensive due to the high costs involving traffic monitoring systems. Another important flaw of these automatic techniques is that due to its cost, only specific critical road segments tend to be monitored e.g., segments with traffic jams or road segments with high percentages of traffic accidents.

In this paper, we explore novel automatic techniques to approximate traffic counts that might alleviate the limitations surrounding both manual and traditional automatic techniques: costs and scalability. This is specially true for emerging economies with limited resources such as Senegal that are looking for alternative ways to measure traffic at large-scale and in an affordable manner. Specifically, we propose the use of Call Detail Records (CDRs) to evaluate traffic volumes on roads. Our proposal is based on the fact that cell phones have become a pervasive sensor of human behavior due to its penetration rates. In Senegal, cell phone penetration rates are well over 94% and the telecommunications infrastructure covers all the country with different levels of granularity mostly between rural and urban environments. Since this infrastructure collects information about the location of the cell phones, we propose to this information as a proxy for traffic count estimation.

The exists an important body of work in the general area of traffic estimation using different sources of sensing data including cameras [9], cell phone data (mostly handover information) [5] or passive data [10] [14]. However, these approaches typically focus on identifying a plethora of traffic issues including road congestion or traffic routes. In this paper, we propose an approach to a problem that although simple, would be very helpful for developing countries with limited resources and for whom traffic counts are a very relevant measure. Additionally, this approach opens the door for the development of cost-effective measures to estimate road conditions in areas where a cell phone network is deployed but no infrastructure to estimate traffic has or can be developed. The rest of the paper is organized as follows: section 2 explains the type of datasets used in our project; section 3 describes the methodology we propose to approximate traffic counts from CDR data and section 4 presents our results. We finalize covering related work and drawing our conclusions in sections 5 and 6.

2. DATASETS

2.1 Traffic Counts

We make use of the latest collection of official traffic counts in Senegal computed in 2002 by national authorities with support from the African Development Bank [2][1]. The official traffic counts contain information for four types of roads: national, regional, departmental and provincial (see Table 1). National roads communicate large administrative regions (*Régions*) and neighboring states and around 60% of them are paved; regional roads connect the fourteen *Régions* in Senegal; departmental roads communicate departments (*Départements*) and around 18% of them are paved; and finally provincial roads connect all other minor urban and production centers. Figure 1 shows a map with the National roads and their official traffic counts (color-coded).

Road Type	Network Length
National	3351km, 59.2% paved
Regional	1206km, 12.6% paved
Departmental	5667, 18.4% paved
Provincial and others	4450, 5% paved

Table 1: Types of roads and lengths in Senegal.

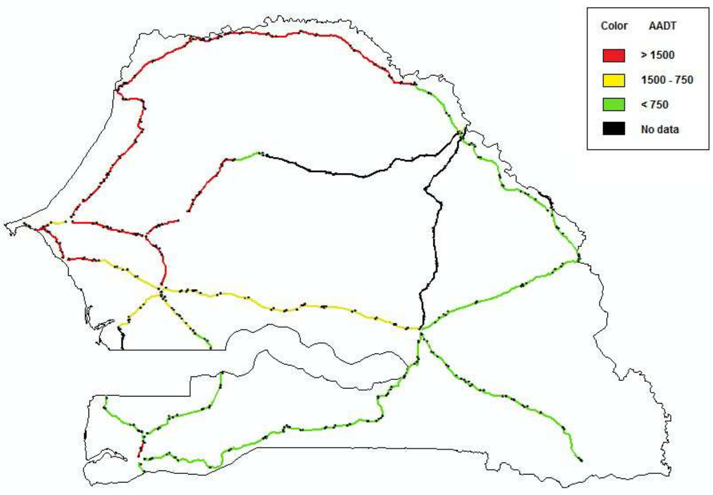

Figure 1: Map of National roads with traffic counts. Green color implies low counts, yellow middle range and red high traffic counts (black is used when counts are not available).

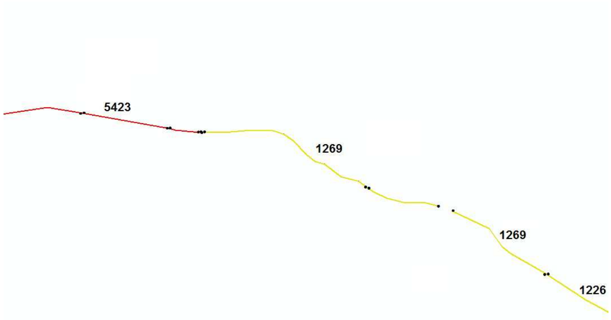

Figure 2: Partial view of National one with several road segments and their traffic counts.

To compute the official traffic counts, each road is divided into a set of segments (links) and traffic counts are provided for each of the individual segments. Each traffic count represents the annual average daily traffic (AADT) or average number of daily cars observed in any given road segment throughout a year. Figure 2 presents a partial view of several road segments in National one showing high and medium traffic count values.

2.2 Call Detail Records

Call Detail Records (CDRs) are collected by telecommunication companies for billing purposes. Every time a phone call is made or received, a set of variables are saved including the anonymized cell phone numbers, date and time as well as information regarding the latitude and longitude of the cellular tower that handled the service. The dataset used for this challenge contains CDRs from Senegal collected for a year: from January 1st to December 31st, 2013. In this project, we make use of *Dataset 2* which contains fine-grained mobility data on a rolling 2-week basis at an individual level with $300,000$ randomly sampled users whose cellular activity spans at least 75% days of the whole year and whose interactions are kept under $1,000$ per week.

Figure 3 shows the map of Senegal with the Voronoi polygons approximating cellular coverage areas in different parts of the country. As observed in other countries [8] [12], the density of towers increases in urban areas i.e., the coverage areas are much larger and less granular as we move from urban to rural areas. It is important to highlight that the real position of the cellular towers is unknown since noise has been added due to commercial and privacy concerns. However, the noise added is proportional to the density of cellular towers and preserves the overall structure of the mesh. Next section also covers how this limitation might affect our methodology. Finally, combining both traffic counts and CDR data, the coverage area of a cellular tower can be traversed by one or multiple road segments as seen in Figure 5.

3. METHODOLOGY

In this paper, we propose to use the cellular activity observed in any given tower as a proxy to predict the traffic that goes through the road segments that traverse the tower's coverage area. In other words, we use the cellular activity computed from the CDR data in *Dataset 2* to predict the official traffic counts per road segment collected by the African Development Bank. However, the cellular activity might be due to individuals who are walking by the coverage area or to individuals who are driving by. Since traffic counts report the daily average number of cars observed in a road segment, we need to envision techniques that compute the cellular activity exclusively as a measure of the motorized traffic. In remote areas of Senegal, located far away from cities and towns, it is highly probable that the cellular activity will be mostly due to individuals driving by the cellular towers. However, in urban environments both types of motorized and non-motorized traffic will take place and will be captured as cellular activity. To disentangle the non-motorized traffic from the overall cellular activity we propose two different approaches: Filter Regions and Filter Users.

3.1 Filter Regions

This approach only considers the cellular activity at towers that give coverage to geographical areas that are mostly inhabited. The idea is that by focusing on rural and remote areas of Senegal with very little residential life with respect to its total geographical area, the cellular activity will mostly represent individuals driving on road segments that traverse the coverage area of the cellular tower. For that purpose, we eliminate from our analysis all coverage areas that are located within and nearby towns with populations larger

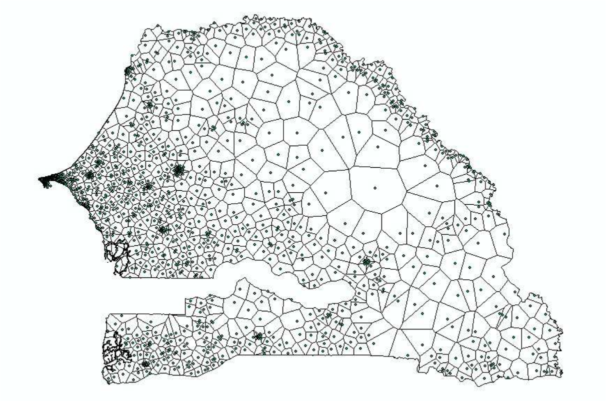

Figure 3: Map of Cellular Coverage in Senegal with Voronoi Polygons.

than $1,000$. We have observed that towns with populations smaller than $1,000$ are typically covered by a cellular tower that spans a geographical area much larger than the towns themselves. Thus, the amount of people possibly walking by remains very small with respect to the total activity observed for the coverage area, making our approach sound for the identification of motorized traffic. Figure 4 shows an example of the Filter Regions approach. Each circle surrounding a population represents the regions (coverage areas) that we filter from our analysis. The circle size is proportional to the population of the town. With this approach, only the coverage areas not filtered are used to approximate the official traffic counts per road segment. Finally, it is important to clarify that this approach offers the advantage that only aggregate information per cellular tower is used without the need to access individual location patterns from users. However, this approach is only useful for computing traffic counts in roads that cover regions that are mostly inhabited. For this filter, we define the cellular activity in a coverage area i as the average daily number of calls detected in that tower ($ActivityCA_i$).

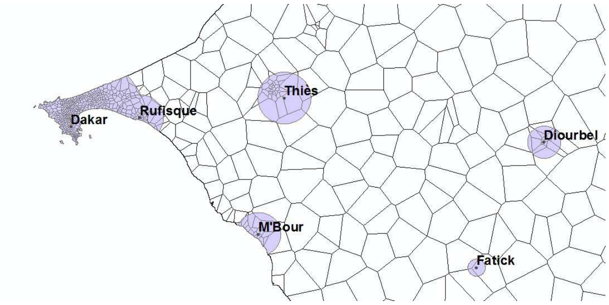

Figure 4: Coverage areas eliminated from analysis with the Filter Regions approach.

3.2 Filter Users

Instead of filtering out highly populated areas, this second approach focuses on filtering out users that are not driving. To accomplish this, we compute the daily distances traveled by each individual and the time it takes them to cover those distances. With these two values in hand, we can compute speed and filter out users whose speeds are lower than a given value that clearly differentiates walking from driving. In this paper, we fix that value at a safe $10mph$ since humans have been reported to be able to walk at maximum speeds of $5.6mph$ [11]. Formally, we can express this second approach as:

$$s_{i,d} = \frac{D(t_j, t_k)}{time}$$

$$driving_d = \{i, with\, s_i \geq 10mph\}$$

where $s_{i,d}$ represents the speed of individual i for a given day d, $D(t_j, t_k)$ the distance between the two farthest away towers for that user in day d and $time$ the time it took the user to cover that distance; $driving_d$ represents the set of individuals i in *Dataset 2* that were driving in day d. In this second approach, we define cellular activity in a coverage area i as the daily average number of users driving through that area ($ActivityCA_i$). As a result, the approach requires the access to individual CDR data but has the advantage of being applicable to any urban or rural environment.

3.3 Predictive Models

Our goal is to explore predictive models that allow us to use the cellular activity of a coverage area $ActivityCA_i$ as a proxy for road segment traffic counts. However, coverage areas and road segments do not necessarily cover exact geographical regions. As a result, we cannot simply assign the cellular activity of a coverage area to a road segment that traverses it. Figure 5 shows the representation of part of National 3 with a set of road segments that traverse the coverage areas of several cellular towers (represented by their Voronoi polygons). We observe that different scenarios can occur: part of a road segment traverses the full coverage area, a complete road segment traverses it, or multiple road segments traverse a unique coverage area. To account for all these scenarios, we approximate the cellular activity for a given road segment j, $ActivityRS_j$, as the weighted average of the cellular activity of all the coverage areas that are traversed by the road segment, weighted by the length of the road segment within each coverage area:

$$ActivityRS_j = \sum_{i=1}^{n} ActivityCA_i * l_{j,i}$$

where $ActivityRS_j$ represents the cellular activity for road segment j, n the total number of coverage areas traversed by the road segment, $ActivityCA_i$ is the cellular activity in coverage area i computed from CDR data and $l_{i,j}$ is the percentage of the length of the road segment j that traverses coverage area i. This is based on the assumption that the cellular activity of a road segment will depend on the cellular activity of the coverage area it traverses weighted by the length of the road segment in that area *i.e.,* the longer the segment, the more activity from the coverage area it captures. It is important to recall that $ActivityCA_i$ can represent the daily average number of calls or the daily average number of drivers in the coverage area depending on whether we use the Filter Regions or Filter Users approach.

At this point we have, for each road segment in our dataset, a ground truth value with the official traffic counts as well as the approximate cellular activity for that road segment, $ActivityRS_j$. By combining together these two values, we frame the prediction problem as a supervised problem with the official traffic counts as the dependent variable and the cellular activity per road segment as the independent variable. For this paper, we explore two types of predictive approaches: linear regression models and Support Vector Regressions. We describe each approach in detail. Formally, the linear regression seeks a model such that:

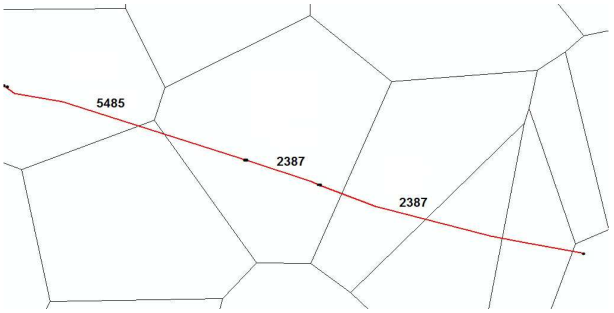

Figure 5: Road segments from National 3 crossing a set of Voronoi Polygons.

$$\hat{tc_i} = \alpha + \beta_1 * ActivityRS_i + \varepsilon$$

where $\hat{tc_i}$ are the predicted traffic counts for road segment i, α and β_1 are parameters of the linear regression and $ActivityRS_i$ is the cellular activity for that road segment. As explained earlier, that cellular activity can be computer using either the Filter Regions or the Filter Users approach. We fit this model with the Ordinary Least Squares approach.

In an attempt to evaluate non-linear models, the second predictive model we explore is Support Vector Regressions (SVR) with a Radial Basis Function (RBF) kernel. SVR with RBF first maps the input to a higher dimensional feature space using a nonlinear mapping (kernel) and then a linear model is constructed in that feature space. Formally,

$$\hat{tc_i} = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

$$K(x_i, x_j) = exp(\frac{\| x_i - x_j \|^2}{2\sigma^2})$$

where K is the kernel function *i.e.,* an RBF in our case and x represents the cellular activity for road segment i ($ActivityRS_i$). Similarly to Support Vector Machines (SVM), SVR looks for the hyperplanes that maximally separate the samples in the projected space although a margin of error tolerance is accepted. SVR uses parameters cost (C) and ϵ to apply a penalty to the optimization for points which are not correctly predicted. In our experiments, we use a grid search to look for the optimal values. Finally, given that we use regressions, we describe the performance of both predictive approaches using the correlation coefficients between the predicted $\hat{tc_i}$ and the official tc_i traffic counts. The resulting correlation coefficient will measure the similarity between the real official counts and the predicted counts using the proposed predictive approaches. Next section describes results using the two predictors and the two filtering approaches.

3.4 Limitations

As discussed in Section 2, the location of the cell towers in the dataset is not the real one. Some noise has been added to avoid publishing the real locations due to commercial and privacy issues. This might affect our methodology since coverage areas are computed based on the cellular tower locations provided by the challenge organizers. If the cellular towers are noisy, the coverage areas will not necessarily represent the real coverage areas and traffic counts associated to those might be noisy as well. However, as reported in the official challenge paper [6], the noise has been added in such a way that the whole mesh structure for the coverage areas is maintained. Thus, although traffic counts associated to the coverage areas might not be the real ones, overall they do represent general traffic trends across different geographical areas. These trends are the ones that we want the prediction models to capture. Although not ideal, the models will still be able to cope with and distil general trends from the noisy cellular tower locations. Similarly, as stated in [6] not all individuals are reported in *Dataset 2* but rather only those with certain lower and upper bounds of activity. Although this could also be seen as a limitation, it is important to highlight that the filter is common across all coverage areas, thus again allowing us to preserve the general trends in our predictive models. Finally, a common limitation in projects that use datasets from different sources is the difference in data collection years: the official traffic counts were collected in 2002 whereas the cellular data is from 2014. Ideally, both datasets should be collected during the same time period. However, the results we report are quite promising despite the time difference in data collection.

4. RESULTS

In this section, we report the accuracy of using cellular activity as a proxy for official traffic counts. Specifically, we describe results for the two measures of cellular activity: Filter Regions and Filter Users and the two types of regression: linear and SVR.

To test the accuracy of the prediction algorithms, we divide each dataset into randomly selected training and testing subsets with a distribution of $80\% - 20\%$ and repeat the random selection 100 times. For each prediction technique and filtering approach, we report average performance results over all 100 runs. Specifically, we report two measures: the correlation coefficient r and the root mean squared error (RMSE). Correlation coefficients report the correlation between the predicted and the official traffic counts whereas the RMSE reports the mean error between the two. In the case of SVR, we use a grid search approach to tune the values for epsilon (ϵ) and cost (C) that give best performance in terms of prediction results. The grid search uses a 10-fold cross-validation over the training set to approximate best values.

Table 2 shows prediction performance for both techniques and filtering approaches during training and testing. Parameter values for SVR are $C = 512$ and $\epsilon = 0.3$ for the Filter Regions approach and $C = 8$, $\epsilon = 0.4$ for Filter Users. First, we observe that SVR obtains better predictive results than Linear Regressions for both types of filtering approaches ($r = 0.698$ vs. $r = 0.512$ in the best case). This result probably reveals that a non-linear approach

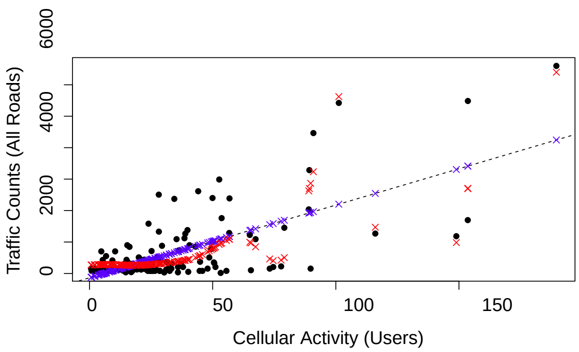

Figure 6: Linear and SVR Regression for one random training run using the Filter Regions approach.

Technique	**Filter Regions**		**Filter Users**	
	Train	Test	Train	Test
Linear Regression	(*r*=0.434,*RMSE*=1676)	(*r*=0.402,*RMSE*=1701)	*(r=0.588,RMSE=1391)*	(*r*= 0.512,*RMSE*=1427)
SVR	(*r*=0.684,*RMSE*=1221)	(*r*=0.572,*RMSE*=1376)	(*r*=0.704,*RMSE*=915)	(*r*=**0.698**,*RMSE*=1218)

Table 2: Prediction correlation coefficients r and RMSE for linear regression and SVR: All Roads.

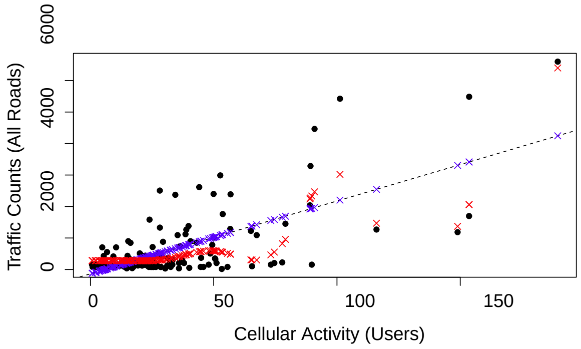

Figure 7: Linear and SVR Regression for one random training run using Filter Users approach.

provides a better fit for the official traffic counts that we want to approximate. To further understand this, Figures 6 and 7 show the official traffic counts (x-axis) and the predicted values during training (y-axis) using linear regression and SVR with the Filter Regions and Filter Users approaches, respectively. As can be seen, the SVR (red crosses in the plot) does a better job at approximating the training samples for both filtering approaches, which also translates into better prediction results during testing as observed in Table 2. The correlation coefficients during testing decrease a little bit probably due to overfitting during training (from $r = 0.704$ to $r = 0.698$).

A second important observation is that the Filter Users approach shows consistently better results than the Filter Regions approach when compared across regression techniques. In fact, the best predictive results were observed for the Filter Users approach and SVR with a correlation coefficient between official and predicted counts of $r = 0.698$ during testing ($r = 0.704$ for training). However, it is also fair to say that the accuracy for SVR and Filter Regions approach had a correct performance with a correlation coefficient of $r = 0.572$ during testing. From these results, it appears that filtering users based on their speed is a more robust approach than just filtering out road segments in urban-like environments.

In an attempt to improve our results, we explored the set of roads for which the traffic count predictions were the worst. We observed that both predictive approaches were doing a poor job with some Regional, Departmental and Provincial roads. As explained in Table 1 these roads are mostly unpaved and the traffic conditions tend to be really poor. Thus, we recomputed the linear regression and SVR for both filtering approaches but only considering National roads. Table 3 shows the prediction accuracy across all scenarios. In general, we observe similar trends to our analysis with all roads: (1) SVR shows more predictive power than Linear Regression for both filtering approaches and (2) Filter Users is more predictive than the Filter Regions approach. However, the main change is that the correlation coefficients between the official and predicted traffic counts were highly improved. In fact, the most accurate results are reported for SVR and the Filter Users approach with correlations of $r = 0.905$ during training and $r = 0.889$ during testing.

Overall, our results indicate that traffic counts can be predicted from cellular activity with good accuracy for analysis including all types of roads and with high accuracy for analysis that only focus on National Roads.

5. RELATED WORK

The exists a wide range of papers exploring the identification of traffic issues including road congestion or traffic routes using different sources of information such as cameras or GPS information [9][13]. Focusing on cell phone infrastructures (GSM and UMTS), Bar-Gera used cell phone data to measure travel speeds and travel times in Israel [3]. The author compared cell phone-based measures with those obtained through dual magnetic loop detectors and showed that there were significant correlations between the two. A similar approach was presented by Chiang *et al.* to extract traffic information exploiting handover patterns from UMTS signals [5]. Moving on to traffic congestion, Janecek *et al.* used data from a cellular mobile network to compute road congestion [10]. They validated their approach against four different traffic monitoring datasets and showed that their methods can detect traffic congestion very accurately and in a timely manner. Other interesting work has also been done in the area of route classification [4]. Becker *et al.* showed how to use handoff patterns from cellular phone networks to identify the routes that people take through a city. The authors presented two classification algorithms to match cellular handoff patterns to routes and validated their methods using statistics provided by a state transportation authority. A similar approach to ours was presented in [7], however the authors used passive network data instead of simply CDRs, making our approach much more challenging. Overall, these approaches typically focus on identifying a plethora of traffic issues in developed countries with plenty of CDR information across customers. Our work aims to develop simple mechanisms with high social impact to be able to approximate travel counts in emerging regions with few resources.

6. CONCLUSIONS

In this paper, we have presented a technique to automatically approximate official traffic counts using mobility features extracted from Call Detail Records. We have evaluated two approaches to detect motorized traffic versus individuals walking: Filter Regions and Filter Users. The former focuses on eliminating highly populated areas where it might be harder to differentiate walking from driving whereas the latter directly eliminates individuals who are not driving. Our results show that SVR together with the Filter Users approach are highly predictive of the traffic counts with correlation coefficients between real and predicted of $r = 0.889$ during testing. As a result, we believe that this approach provides a reliable technique to measure traffic counts at large-scale and in affordable manner.

Technique	**Filter Regions**		**Filter Users**	
	Train	Test	Train	Test
Linear Regression	(*r*=0.604,*RMSE*=1243)	(*r*=0.584,*RMSE*=1321)	(*r*=0.727,*RMSE*=825)	(*r*=0.705,*RMSE*=953)
SVR	(*r*=0.865,*RMSE*=803)	(*r*=0.748,*RMSE*=906)	(*r*=0.905,*RMSE*=610)	(*r*=**0.889**,*RMSE*=682)

Table 3: Prediction correlation coefficients r and RMSE for linear regression and SVR: Only National Roads.

7. REFERENCES

[1] http://dlca.logcluster.org/display/public/dlca/lca+homepage.

[2] http://www.infrastructureafrica.org/documents/type/arcgis-shape-files/senegal.

[3] H. Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, 2007.

[4] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. Route classification using cellular handoff patterns. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11. ACM, 2011.

[5] C.-Y. Chiang, J.-Y. Chuang, J.-K. Chen, C.-C. Hung, W.-H. Chen, and K.-R. Lo. Estimating instant traffic information by identifying handover patterns of umts signals. In *IEEE International Conference on Intelligence Transportation Systems*, pages 390–395, 2011.

[6] Y. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *CoRR*, abs/1407.4885, 2014.

[7] V. Frias-Martinez, Y. Moumni, and E. Frias-Martinez. Estimation of traffic flow using passive cell-phone data. In *Workshop on Data Science and Macro-Modeling (DSMM), SIGMOD*, 2014.

[8] V. Frias-Martinez and J. Virseda. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, ICTD '12, pages 76–84, 2012.

[9] V. Jain, A. Sharma, and L. Subramanian. Road traffic congestion in the developing world. In *Proceedings of the 2Nd ACM Symposium on Computing for Development*, ACM DEV '12, pages 11:1–11:10, 2012.

[10] A. Janecek, K. A. Hummel, D. Valerio, F. Ricciato, and H. Hlavacs. Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 361–370. ACM, 2012.

[11] A. Minetti. The three modes of terrestrial locomotion. *Biomechanics and Biology of Movement. Human Kinetics*, pages 67–78, 2000.

[12] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, and N. Oliver. Human mobility in advanced and developing economies: A comparative analysis. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

[13] W. Shen, Y. Kamarianakis, L. Wynter, J. He, Q. He, R. Lawrence, and G. Swirszcz. Traffic velocity prediction using gps data: Ieee icdm contest task 3 report. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1369–1371, 2010.

[14] D. Valerio and et al. Road traffic estimation from cellular network monitoring: a hands-on investigation. In *Personal, Indoor and Mobile Radio Communications, 2009 IEEE 20th International Symposium on*, pages 3035–3039. IEEE, 2009.