Predicting Location of Emergency Response Centers

Baiting Luo, Rishav Sen
Institute for Software Integrated Systems, Vanderbilt University

Baiting Luo]1 Rishav Sen]2 ***Abstract*—Over the years, traffic levels have constantly increased, leading to a larger number of collisions and accidents taking place on the roads. These accidents need to be handled by emergency medical responders as soon as possible. With an increase in the number of accidents taking place, it is necessary that every one of them are addressed with an increased urgency. Although the number of emergency workers are fixed, they still need to provide for and rush to all the accidents happening in and around a large geographic landscape - in our case, Davidson county of Tennessee, in the United States of America. This gives rise to the fact that with a limited number of workers available, we need to make sure that their deployment be made to places that are optimally located to the serve the accidents taking place all across. The classical models that are used to predict the emergency response locations the have been known to undermine the sparsity of data that comes with handling accidents spread over large areas. We provide a framework to address this issue by spatio-temporally aggregating the traffic and the accident data over the area of interest and make optimal predictions for locating the emergency response hubs.**

I. Introduction

Travelling is an integral part of our lives and one the most common modes for is using cars. With a rise in the number of car use, there has a been a surge in the motor vehicle accidents. Road accidents alone account for 1.25 million deaths globally and about 240 million emergency medical service (EMS) calls are made in the U.S. each year [1]. Safety and efficiency are the two primary goals of transportation engineering. The effort that public agencies put into reducing traffic accidents is highly justifiable. They work to make better plans to reduce the impact of such incidents by allocating more resources, making better deployment decisions as traffic accidents place a huge financial burden on society. It comes as a moral and binding requirement for all government agencies to treat this issue with high priority. It is through the formation for Emergency response management (ERM) systems that the vast incidents are handled. It is defined as the set of procedures and tools that first responders use to deal with incidents such as road accidents. This includes specific mechanisms to forecast incidents, detect incidents, allocate resources like ambulances, dispatch resources, and finally mitigate the post-effects of incidents [1]. But, these systems come at a cost. Increasing the number of first responders is not always an option as many agencies face shortage of funds. The lack of proper road-side infrastructure is also a possible shortcoming as responders cannot be deployed very frequently along the roads. Such situations demand that the few number of ERM facilites that can be maintained properly, be placed at the best possible locations in order cater to the traffic accidents happening in ad around the place. Also, the need to be divided into zones. Else, their resources could get wasted on travelling to and from the accident location, and spending more time than required. Along with these constraints, we also need to take into account that the accident rates are not the same for all the roads throughout the year. They may vary depending on the weather and traffic conditions [2]. The temporally varying trends affect the placement of any ERM facility. With varying traffic conditions, the accident hotspots may change drastically - leaving the first responders stranded with multiple incidents to attend at a time. To avoid such a scenario, both the spatial and temporal trends need to considered in designing the system.

Our model caters to handle such situations by predicting the traffic trends and optimizing the ERM facilities accordingly. It does this by taking into consideration the evolving vehicular patterns over time and finding the regions that are more prone to traffic accidents, while not completely disregarding the less accident areas. The method we propose here is also an online prediction model, being able to work under changing conditions. Online prediction of incidents is becoming more important as more intelligent transportation system deployments in the United States focus on improving the operations of transportation management facilities. Parsing through very large sets of traffic and accident data, our system is robust enough to handle massive data loads - in turn enhancing the predictions it produces.

The algorithm used here to provide the best possible locations is the uncapacitated facilities location problem with exactly *'p'* number of facilities, or more commonly referred to as "*p-median* problem". It is solved thorugh a branch and bound solution [3]. The bounds are obtained by solving the Lagrangian relaxation of the p-median problem using the subgradient optimization method. The proposed algorithm is simple, requires small core storage and computational time, and can be used for solving large problems. This enables us to provide the road segments on which the ERM facilties should be placed.

II. Related Work

The work we discuss here focuses on how to predict the location of accidents and further use the knowledge obtained to locate emergency response centers optimally. The optimality is achieved by placing the response centers in a manner such that the they are as close as possible to the to the accident hotspots. This idea is strongly correlated by Vazirizade *et al*

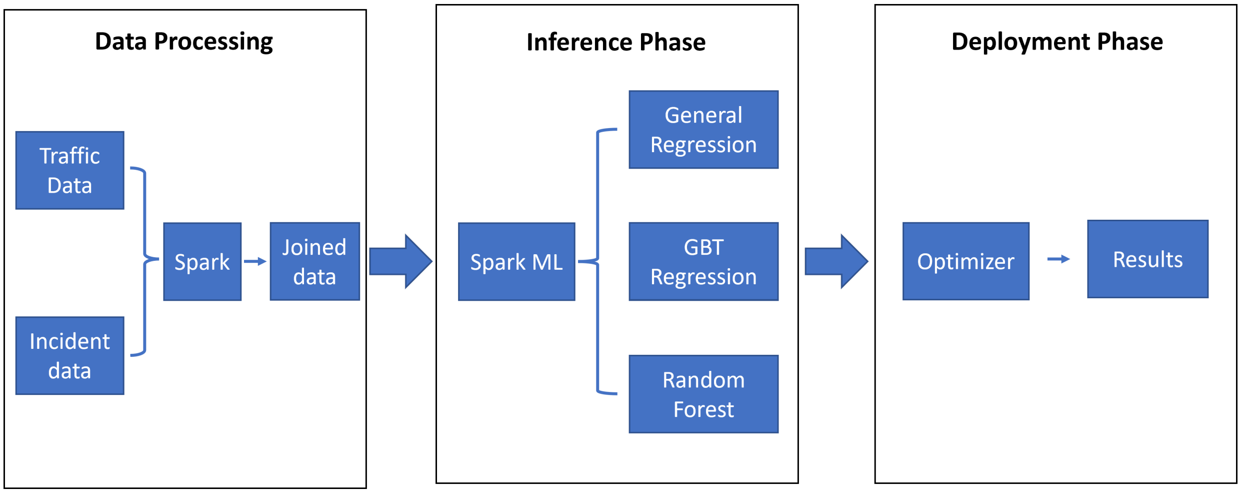

Fig. 1: Workflow

[4]. There has been much work on trying to predict and prevent incidents on major road networks [5] [6]

This study presents an optimization strategy for local agencies to distribute incident response units effectively along freeway segments plagued by frequent incidents. The proposed method, based on the p-median model [3]. It is applied to the road network in Davidson county, Tennessee.

III. Approach

In this project, our approach can be discussed in three phases as shown in Figure 1. In the first phase, we leverage the buffer technique to overcome the difficulty of mapping accident data to traffic data due to the discrepancies of spatial information among different datasets.In the second phase, we leverage the convenience of spark ML to train several different machine learning models given the training and testing datasets obtained from the first phase. In the last phase, we implement *p-median* algorithm based on the accidents prediction results to decide the best locations for deploying emergency response facilities. In this section, each of the three phases is elaborated and discussed.

A. Data Preparation

In the real world, even though the location of accidents can be approximately recorded and located, geospatial data is by nature uncertain and often incomplete due to being abstractions and observations of a continuous reality [7]. Moreover, lots of spatial analysis is done at a pointwise scale, which further increases difficulty of data mining, knowledge discovery when multiple datasets are counted in. Therefore, certain data mapping technique needs to be introduced to address the data mapping issue caused by discrepancies of spatial information among different datasets. Assume we are given an accident dataset and a county dataset:

$$D_{accident} = \{(l_1, t_1), (l_2, t_2), ..., (l_n, t_n)\} \tag{1}$$

$$D_{county} = \{(s_1, g_1), (s_2, g_2), ..., (s_n, g_n)\} \tag{2}$$

where l_i indicate the pointwise location of the accident with unique ID=i, t_i indicates the incident time for accident i, s_m indicates the road segment with a unique ID=m, and g_m indicates the geometry information for road segment s_m.

Ideally, we can join two datasets $(D_{accident} \bowtie D_{county} \mid l_i = g_m)$ to map the accident to the corresponding road segment. However, such straightforward method is normally not approachable due to the spatial information's discrepancies among datasets.

Therefore, we leverage a buffer strategy which converts the incident point into a search area in the shape of circle. Therefore, for each of accidents in $D_{accident}$, it has one more attribute sa to indicate the geometry information of the search area. Hence, for each accident, we are trying to find a road segment which is the closet to it:

$$(s_x, g_x) = \underset{(s,g)\in D_{county}}{\text{argmin}} \quad (Distance(g, l) \mid s \cap sa) \tag{3}$$

, in which $Distance$ and $s \cap sa$ can be done via functions of GeoPandas[1]. After these processes, a new accident dataset $D^*_{accident} = \{(l_1, t_1, s_1), (l_2, t_2, s_2), ..., (l_n, t_n, s_n)\}$, which is merged with road segment ID, can be obtained. Different from accident data, of which geometry information is pointwise, traffic dataset is always clearly labeled with road segment ID and timestamp. Therefore, given a traffic dataset:

$$\begin{aligned} D_{traffic} = \{&(s_1, t_1, traf_1), (s_2, t_2, traf_2), \\ &..., (s_n, t_n, traf_n)\}. \end{aligned} \tag{4}$$

, where $traf_n$ indicates the traffic information such as average speed, congestion rate of road segment s_n at time t_n. Then a dataset ready for training can be easily obtained by:

$$\begin{aligned} D_{training} = (&D^*_{accident} \bowtie D_{traffic} \mid \\ &(s, t) \in D_{accident} = (s, t) \in D_{traffic}) \end{aligned} \tag{5}$$

Overall, as shown in Figure 2, the workflow of data preparation can be described as generating $D^*_{accident}$ first, then merging $D_{training}$ and $D^*_{accident}$ with equation 5.

[1] http://geopandas.org, doi:10.5281/zenodo.2585848

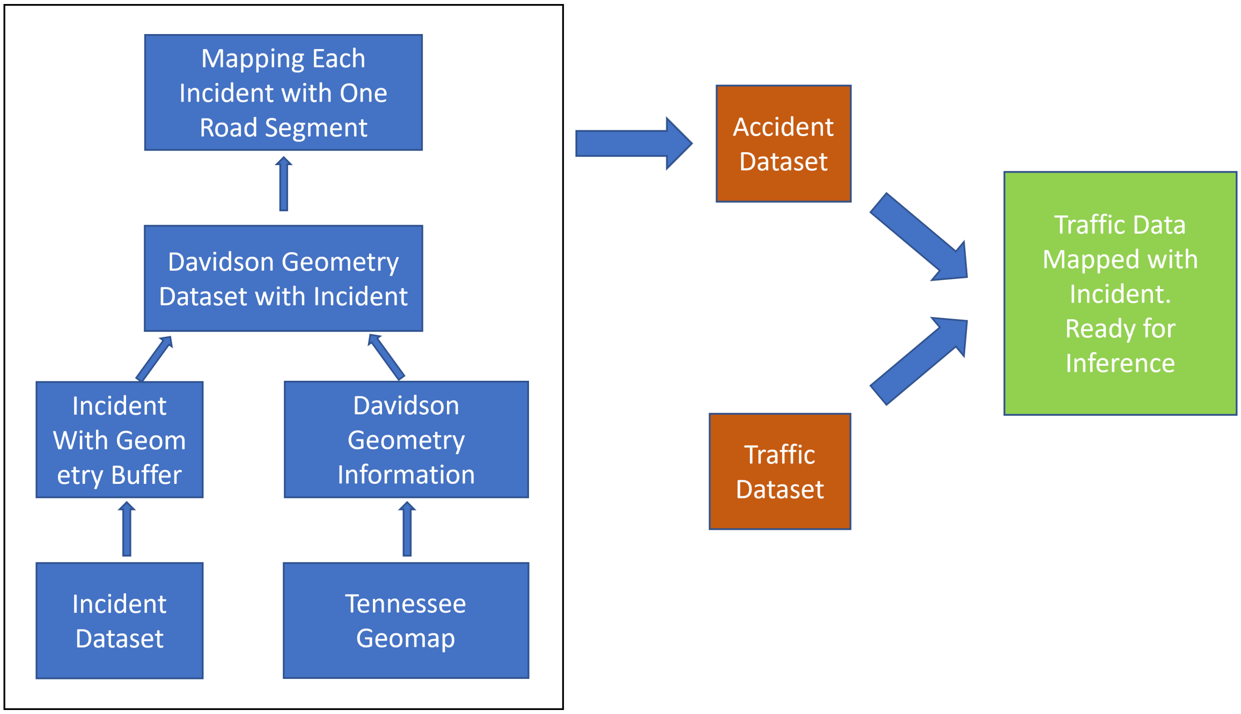
Mapping Each Incident with One Road Segment
Davidson Geometry Dataset with Incident
Incident With Geom etry Buffer
Davidson Geometry Information
Incident Dataset
Tennessee Geomap
Accident Dataset
Traffic Dataset
Traffic Data Mapped with Incident. Ready for Inference

Fig. 2: Data Preparation

B. Machine learning Models

In this work, we train three machine learning models with $D_{training}$ for inferring if there are any accidents in the unseen dataset $D_{testing}$. They are separately poisson regression, random forest, gradient boosted tree regression. In this section, the background knowledge of these three models are introduced.

1) Poisson Regression:

For a poisson distribution, the expected value is λ. Therefore, with the absence of other information, one can expect to see λ events in any time interval t. For any given time interval t, a number of λt events can be expected. However, if λ can change from one observation to the next, we assume λ is influenced by a vector of explanatory variables, namely, regressors. In our dataset, we have a matrix of regression variables $X = traf \in D_{training}$ as regressors and observations $Y = (accident \in 0, 1) \in D_{training}$, in which 0 indicates there is no accident and 1 indicates there is an accident.

Given these variables, the poisson regression model [8] is used to fit the observed counts Y to the regression matrix X via a link-function that expresses the rate vector λ as a function of the regression coefficients β and the regression matrix X.

2) Random Forest:

Random forest [9] (RF) works as its name implies, leveraging the ensemble of a large number of individual decision trees to decision making. Each individual tree in the RF spits out a class prediction, and the class with the most votes become the model predictions. More specifically, RF is a meta estimator that fits a number of trees on various sub-samples of the dataset, leveraging average values of sub-samples to prevents overfitting and improve prediction accuracy.

Different from general linear regression models, the key point of RF is that there is a low correlation between sampled trees. Each of these sampled trees can protect each other from its individual error. To be specific, even if some trees are wrong, as long as the majority of trees is correct, the group of trees can still move towards correct direction.

To ensure such low correlation between trees, a bagging [10] (bootstrap aggregation) technique is leveraged. RF takes advantage of bagging by making each individual tree to randomly sample data from dataset with replacement. Furthermore, each tree in a random forest is forced to pick from a random subset of features; in this case, more variation is created among the trees and leading to a lower correlation across the trees.

3) Gradient Boosted Trees:

As a member of regression tree family, GBT [11] regression model is mainly different from RF model in how the decision trees are created and aggregated. More specifically, each tree in GBT is built after another. The concept of boosting comes from the idea of using new tree to improve on the deficiencies of the previous trees. Furthermore, the gradient indicates that the algorithm minimizes the gradient of the loss function as the algorithm builds each tree. Another key difference is that RF aggregates the results of decision trees at the end of the process which GBT aggregates the result of each tree along the way to get the final result. Although GBT can empirically obtain better results than RF, it is notorious for easily overfitting. Moreover, due to the differences between RF and GBT described above, GBT theoretically take much longer time to run than RF given a dataset with the same size.

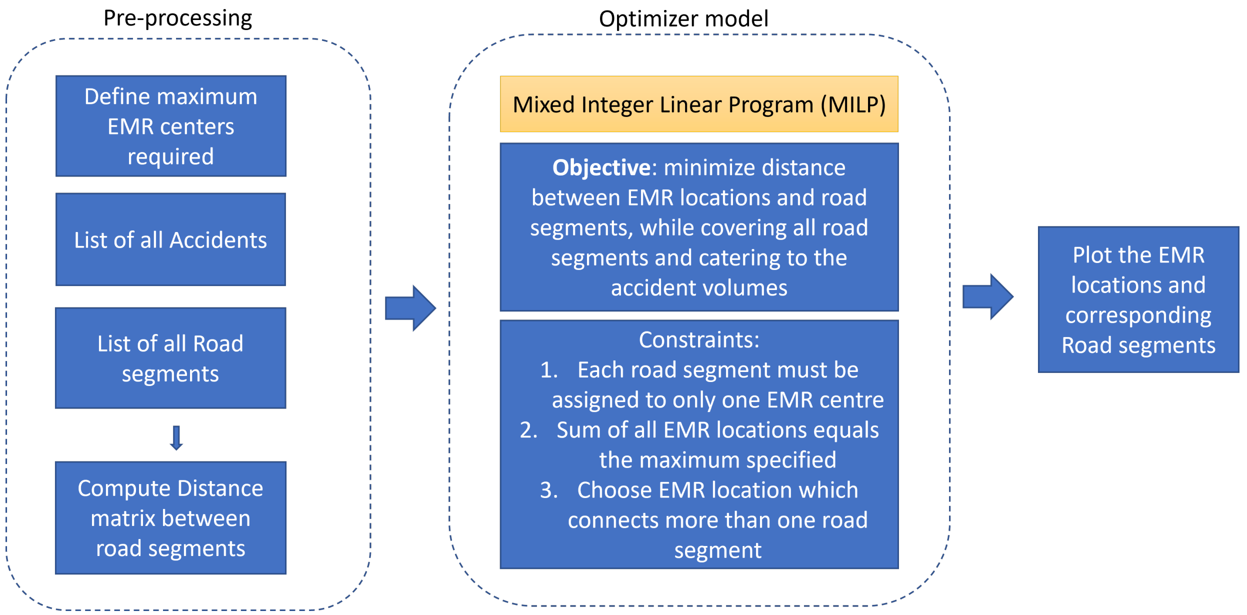
Pre-processing
Define maximum EMR centers required
List of all Accidents
List of all Road segments
Compute Distance matrix between road segments
Optimizer model
Mixed Integer Linear Program (MILP)
Objective: minimize distance between EMR locations and road segments, while covering all road segments and catering to the accident volumes
Constraints:
1. Each road segment must be assigned to only one EMR centre
2. Sum of all EMR locations equals the maximum specified
3. Choose EMR location which connects more than one road segment
Plot the EMR locations and corresponding Road segments

Fig. 3: Data Preparation

C. Optimization Method

After having generated the predictions for the accident rates, we would require to aggregate them over time to perform further analysis. The problem we are aiming to solve is that of EMR facility location using a distance-based aprroach. The optimization algorithm that we look into is the *p-median problem* [3]. It can be solved in polynomial time on a tree structure, i.e., a structure without loops. This algorithm works by locating the central deployment locations and allocate the available resources in proportion to the number of traffic accidents taking place, as shown in 3.

Its mathematical model here consists of the number of EMR facilities denoted by $e \in E$. There are $r \in R$ number of road segments for which we have the accident data and need to decide on. We also define the demand $z \in Z$ to be the number of accidents taking place at each road segment. The demand at any road segment that needs to be met the EMR facility is $z_{r,e}$ $\forall$ $e \in E, r \in R$. The algorithm is also provided with the knowledge of the distances between each of the road segments. This information is needed as it needs to decide the EMR facility location based on the corresponding road segment distances. It should only attempt to cater to the segments close to it, dependent on the number of accidents predicted. The distance matrix is computed for this, by computing the geometric distance between the road segments. This is done by leveraging the distance method in Geopandas [2] and is presented as $d_{e,r} \in D$ $\forall$ $e \in E, r \in R$. Since we always want a predefined value of EMR facilities to be located around the roads, we defined the number of facilities to be P.

Now that the data values defined, we also need to introduce a couple of decision variables introduced to solve the problem at hand,. We choose the decision variables such that both of them are binary. The first is whether to allocate a road segment r to an EMR facility e and is denoted as $A_{r,e} \in \{0, 1\}$. Next, we need to know if the EMR facility will be located at road segment r or not, in the form of $S_r \in \{0, 1\}$. Now we need to formulate our objective such that we can have the least number of EMR locations serving the largest number of road segments possible, without any overlap. This also introduces some constraints on our objective function which are denoted accordingly. The objective function is defined as:

$$\min \sum_{e=1}^{E} \sum_{r=1}^{R} z_r \; d_{e,r} \; a_{e,r} \tag{6}$$

$subject\ to$

$$\sum_{e=1}^{E} A_{e,r} = 1, \;\; \forall \; r \in R \tag{7}$$

$$\sum_{e=1}^{E} S_e = P \tag{8}$$

$$A_{e,r} \leq S_e, \;\; \forall \; e \in E, r \in R \tag{9}$$

The constraints cover the conditions that we would like to enforce on our model. We want each road segment to be assigned to only one EMR facility using 7. The maximum number of EMR facilities should always be equal to the maximum value and is set using equation 8. Equation 9 states that the EMR facility must at least cover one road segment.

IV. Experiment Results

A. Machine Learning Prediction Result

Due to the sparsity of accidents data, the dataset used for training would be extremely unbalanced if all given datasets

[2] http://geopandas.org, doi:10.5281/zenodo.2585848

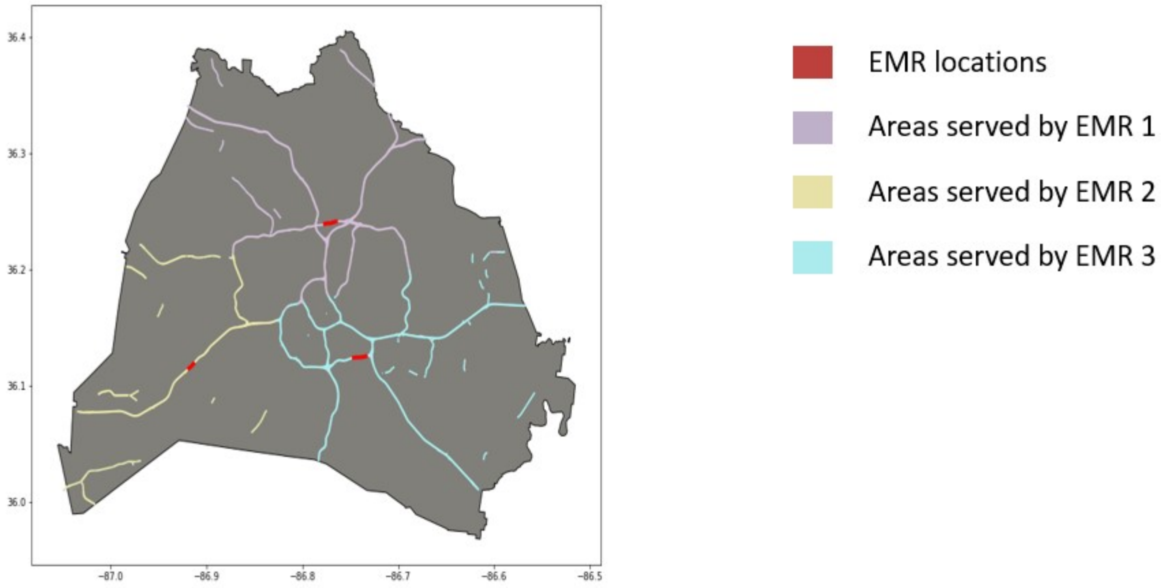

Fig. 4: Optimized EMR facility locations and the areas served by them

are used for training together. To solve this, we only use all traffic data associated with an accident and mixed them with the same amount of traffic data without any accident. In this case, we are able to obtain a balanced dataset for training.

However, an entire unseen dataset by trained ML models should be used for testing the performance of trained ML models. To overcome the issue of unbalanced testing dataset, the results are evaluated with precision, recall and F1 scores [12]:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{12}$$

where TP, FP, FN are respectively true positives, false positives and false negatives.

As shown in table I, random forest and gradient boosted tree models have the similar performance. Both models have good results at recall results, which indicate that both models can correctly classify the traffic data associated with the accident. However, both models have bad performance at recall results, which indicate lots of traffic data without any accident is classified as the traffic data associated accident. As our testing dataset has thousands of traffic data without accident while only a few traffic data with accident, this phenomenon can be explainable. Furthermore, as indicated in Section III-B3 that the running time of GBT would theoretically be much longer than RF, the experiment results further verified this. Compare with GBT and RF, Poisson regression model has the worst performance. This can be explained as Poisson regression makes the assumption that there is a Poisson distribution behind the dataset while our dataset is actually not aligned with this assumption.

TABLE I: Prediction Results Comparison for ML Models

	Precision	Recall	**F1**	Time
Random Forest	0.48	0.97	**0.64**	2 mins
Gradient Boosted Tree	0.51	0.91	**0.65**	20 mins
Poisson Regression	0.41	0.78	**0.54**	1 min

B. *Optimization Result*

The predicted data produces the number of accidents taking place for every road segment. The values of accidents for each road segment are aggregated for the desired interval of time. This gives the sum of accidents that have taken place. The sum of accidents is the demand value for the p-median model. the other required parameters R and E are set as the list of all the road segments, while the distance matrix is calculated for $r \in R$. We also define the maximum number of EMR facilities to be 3 ($P = 3$). Now we can evaluate the result. For the predicted data of February 2022, and from amongst over 8000 road segments, the optimizer provides us the best suited EMR facility locations to be on 3 road segments.

$$e_{12} = 1, e_{193} = 1, e_{378} = 1, \ \forall \ e \in E$$

This data is then plotted on the map of Davidson county, Tennessee, USA in Fig.4. The road segments in red are the proposed EMR facility locations, considering the temporal accident patterns around them. The three different colored road segments are the separate roads to which each of the EMR facilities should take care of. For comparison, we also plot the

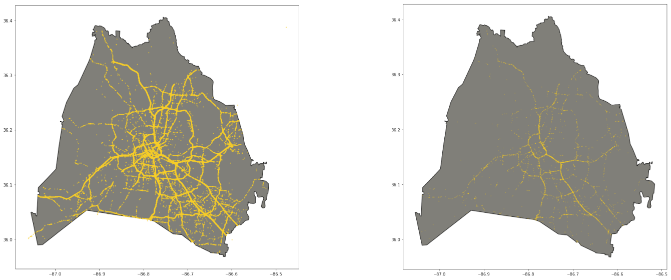

Fig. 5: (left)All accidents from Mar’ 2019 to Jan’ 2022. (right) Ground truth of accidents for Feb’ 2022

map of all accidents, from March 2019 to January 2022 and the ground truth values for the month predicted, i.e., February 2022 in Fig.5

V. Conclusion

The road networks are congested day by day and we need a proper solution to tackle the rising number of traffic collisions. With limited resources, the local governments have their hands tied in trying to provide emergency medial aid to all locations of accidents. The work proposed here is aimed at this shortcoming. We provide a robust method to predict and optimize the placement of EMR facilities across a large geographical area. This can in turn reduce deployment costs and save crucial time in responding to traffic accidents.

References

[1] A. Mukhopadhyay, G. Pettet, S. Vazirizade, Y. Vorobeychik, M. J. Kochenderfer, and A. Dubey, “A review of emergency incident prediction, resource allocation and dispatch models,” *CoRR*, vol. abs/2006.04200, 2020. [Online]. Available: https://arxiv.org/abs/2006.04200

[2] M. G. Karlaftis and I. Golias, “Effects of road geometry and traffic volumes on rural roadway accident rates,” *Accident Analysis & Prevention*, vol. 34, no. 3, pp. 357–365, 2002.

[3] S. C. Narula, U. I. Ogbu, and H. M. Samuelsson, “An algorithm for the p-median problem,” *Operations Research*, vol. 25, no. 4, pp. 709–713, 1977.

[4] S. M. Vazirizade, A. Mukhopadhyay, G. Pettet, S. E. Said, H. Baroud, and A. Dubey, “Learning incident prediction models over large geographical areas for emergency response systems,” *arXiv preprint arXiv:2106.08307*, 2021.

[5] Y. Qi and H. Teng, “An information-based time sequential approach to online incident duration prediction,” *Journal of Intelligent Transportation Systems*, vol. 12, no. 1, pp. 1–12, 2008.

[6] C. A. Miller, “Development of a decision support framework for deployment of incident management systems,” *ProQuest Dissertations and Theses*, p. 327, 1999, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2022-01-10. [Online]. Available: http://proxy.library.vanderbilt.edu/login?url=https://www.proquest.com/dissertations-theses/development-decision-support-framework-deployment/docview/304530310/se-2?accountid=14816

[7] A. U. Frank, “Tiers of ontology and consistency constraints in geographical information systems,” *Int. J. Geogr. Inf. Sci.*, vol. 15, no. 7, pp. 667–678, 2001. [Online]. Available: https://doi.org/10.1080/13658810110061144

[8] E. R. Ziegel, “An introduction to generalized linear models,” *Technometrics*, vol. 44, no. 4, pp. 406–407, 2002. [Online]. Available: https://doi.org/10.1198/tech.2002.s91

[9] “Random forests,” in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Springer, 2010, p. 828. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_695

[10] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: https://doi.org/10.1007/BF00058655

[11] D. Steinberg, M. Golovnya, and N. S. Cardell, “Stochastic gradient boosting: An introduction to treenet™,” in *The 15th Australian Joint Conference on Artificial Intelligence 2002, Proceedings Australasian Data Mining Workshop, Canberra, Australia, 3rd December 2002*, S. J. Simoff, G. J. Williams, and M. Hegland, Eds. University of Technology Sydney, Australia, 2002, pp. 1–12.

[12] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, 1st ed. Springer Publishing Company, Incorporated, 2008.