**CSE564 Final Project Report**

Xinhe Huang 108390641

Silu Ou 110563639

**Problem & background**

Some employees of GAStech disappeared in the last two weeks. Since they used the cars of the company and the data of these cars’ movements as well as the data of the employees’ transactions were both recorded, we can analyze these data to find out the reason of their disappearance.

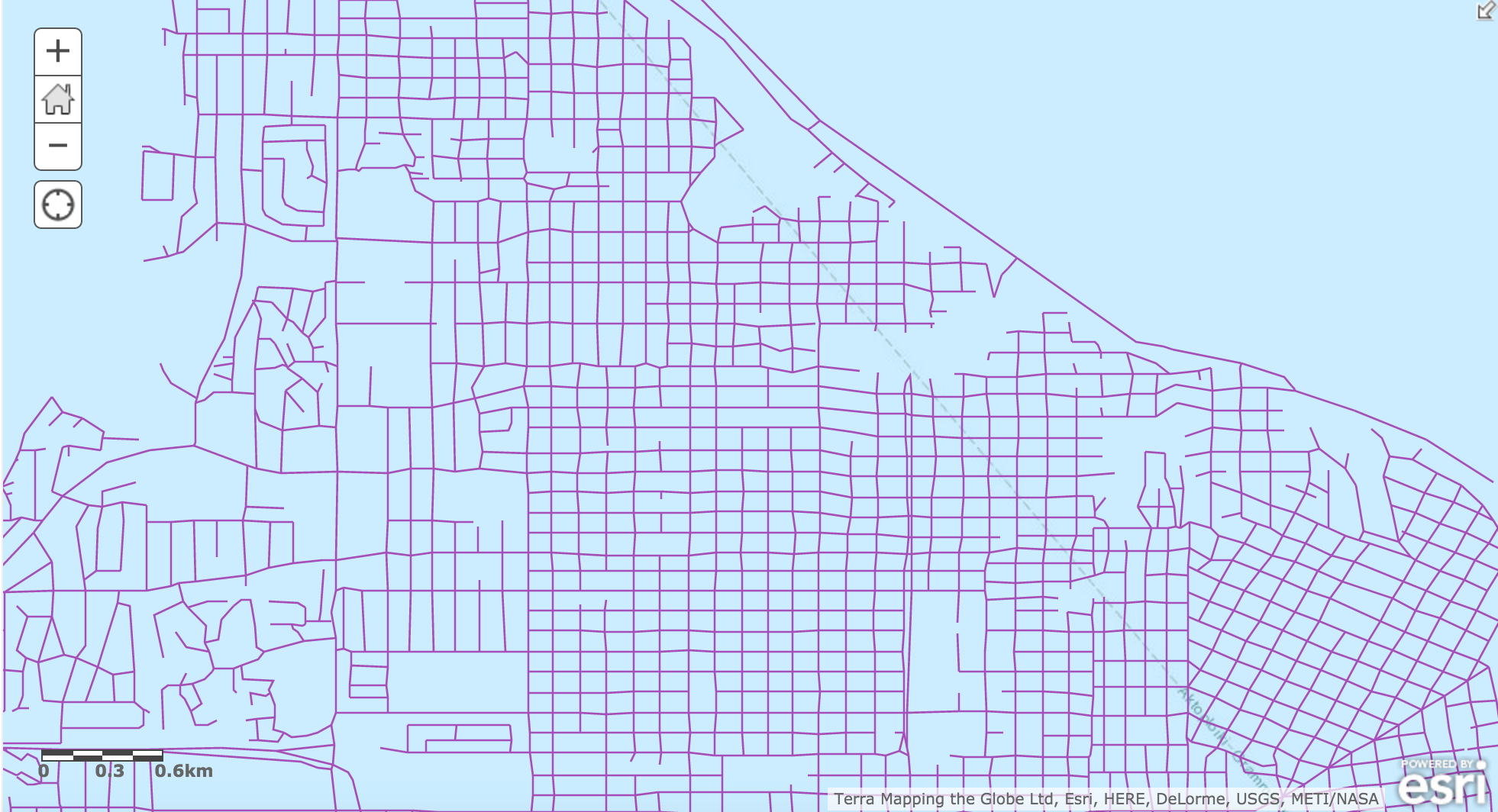
In order to find out why they disappeared, we need to answer the following questions:

1. What does a day in the life of a typical GAStech employee look like?
2. What unusual events or patterns do we observe? For this question we need to find out the location and the person involved as well as the time when an event occurred. We also need to decide if it is necessary to pay attention to this pattern.
3. How to deal with the imperfect data?

**Preparation**

In order to get a better understanding in the *gps.csv* file, we use the app ArcGIS to open the ESRI shapefile *Abila* and translate the map into a file named *location.csv*. Then we use *location.csv* and *gps.csv* to find out the places that each car has visited and get a new file *route.csv* which records the start place and the end place of each car’s routes. This part of work is done with python.

The ESRI shapefile *Abila* and the map are shown as below.



Abila



map

**Data Processing**

Before the prelim report we finished the visualization of the dataset of credit card transactions. In this part we first did the filtering work. We deleted the records of the people that only go to one or two places and have transactions there because these records only have little useful information for data analysis. Then we calculated the frequency of each person’s transaction at each location to find out their unusual behavior. In addition, we calculated the total price of each person’s transaction at each location. Then we compared the total price with the frequency of its corresponding location to see if there is any abnormality.

We also did some pre-analysis work on the gps dataset. First we found that *gps.csv* file records each car’s gps at every second as long as the car is moving. Thus we handled the data in this file to get the start location and the end location of each journey.

After the prelim report we continue the data analysis work on the gps dataset. For each car we calculate the frequency of its visit to each place. Since we can not get the exact gps of each location on the map, we can only attach the gps of each jouney’s end location to the nearest location on the map. If one journey starts and ends at the same location, then it means that the employee who this car is assigned to hangs around this place for this journey. Then we complete the visualization work of this part of dataset.

We also add a line chart for credit card transaction dataset. We decide to show a line chart of all the transactions in each place that an employee visit. This chart is used for further analysis on transactions of each employee. If one person has a transaction of especially high or low price than his/her normal transactions in one place, this activity needs more attention since it seems abnormal.

**Interaction**

In this section we will give examples that how users can interact with the server and get the data they want.

1. Overall Look

Below is the whole web page to show the graphs of all the datasets.



We use a bar chart to show the total price of transactions in different places, a pie chart to show the frequency of places a person visited, a line chart to show the details of transactions and a pie chart to show the frequency of nearest places visited by the car assigned to a person. Besides there is a button on the top left corner for users to choose different people’s data.

2. Choosing different people

If we want to choose different people’s data to do analysis work, we can click the button on the top left corner and then there will be a list of people shown as below:



Then we can click the name of the person we want and the graphs will be updated. For example, if we choose Birgitta Frente, the web page will be updated as below:



3. Showing total price of transactions

When move mouse to bars in the bar chart, the total price of transactions in the corresponding place as well as the name of this place will be show on top of the bar as below:



4. Showing frequency of places visited for transaction

When move mouse to different parts in the right pie chart, the number of visits to the corresponding place will show on top the mouse as below:



5. Showing details of transactions

If we want to get details of transactions in a place, we can click the corresponding parts in the pie chart. Then a line chart will show the price of transaction on each day at this location. When move mouse to the point, the price and the date of this transaction will show on top of the point as below:



6. Showing frequency of nearest places visited by the assigned car

When move mouse to the left pie chart, the number of visits by the assigned car to the corresponding place will show on top the mouse as below:



**Analysis on Charts/Graphs**

In this section we will give some examples that we do analysis on. We choose 3 people to show how we find their unusual behaviors.

**1. Birgitta Frente**



From the charts above, we know that Birgitta’s assigned car often went to places near Gas Tech like others. Since Gas Tech is the company he is working for, this means his activities were mainly in the range of his company. We also notice that there are two places his car visited only once, which are in the beginning of the gps records. As in other cars gps records there are also one or two places visited only once in the beginning, we decide to recognize them as normal behaviors and ignore them.

One abnormal activity in credit card transactions is that Birgitta just made one transaction at Roberts and Sons. But this transaction has the highest price. Maybe it is because Birgitta bought something extremely expensive in Roberts and Sons. But we still recognize it as an abnormality.

Another abnormal behavior in credit card transactions is in Shoppers’ Delight. There is also only one transaction in this place. Although it does not have the highest price, its price is still much higher than other places which only have one transaction.

For the details of transactions, we also find some abnormalities. In Katerina’s Café Birgitta often bought things with price lower than $25. However, there are three transactions having prices higher than $30, one of which is even around $55. Since Katerina’s Café has a high total price of transactions, we recognize these three transactions as abnormal activities.

We also notice that even the car’s gps data show that Birgitta’s activity range is mainly in the company, the credit cards transaction data show that Birgitta actually often went to other places. This means Birgitta did not use the assigned car to these places.

**2. Ada Campo-Corrente**



The assigned car of this guy, Ada Campo-Corrente, has similar records with Birgitta. He also used his car mainly around the company while he also went to other places without the assigned car and had transactions there.

For the credit card transaction data, we find that Ada had only two transactions in Desafio Golf Course but the total price of these two transactions are almost the highest. Then we look into the details of these two transactions. The graphs are shown as below:



It seems that Ada paid the most in the first transaction. We assume that it is because he needed to pay the most fees of the Golf Course before he took the course. After we analyze other people who also had transactions in Desafio Golf Course, we find that these people all had such behaviors and decide to recognize them as normal behaviors.

Ada also had an abnormal transaction in Shoppers’ Delight like Birgitta. There is only one transaction there but the price is much higher than other places only having one transaction except Kronos Mart. Kronos Mart also has only one transaction whose price is almost the same as the one in Shoppers’ Delight. We recognize the transactions in these two places as abnormal behaviors.

For the details of transactions, we find some abnormality in Hippokampos. Most of the transactions in this place have price lower than $50. But there are two transactions having price higher than $60. Also there is a large oscillation in the line chart. The lowest price is $12.47 while the highest is $63.21, which is next to the lowest transaction. We think that these are all abnormal.

**3. Lidelse Dedos**



According to the charts above we can see that the records of the assigned car of Lidelse Dedos mainly went around the company although he also had transactions in other places, just like the records of Birgitta and Ada indicate.

When it comes to the credit card transaction data we notice that Frydos Autosupply n’ More has the highest total price of transactions although Lidelse only went there three times. Thus we look into the details of these three transactions as below:



From the graph we can see that the third transaction has a much higher price than other two in this place. Since we are not very sure what this place sell, we decide to recognize the third transaction as an abnormal behavior.

We also find that for Katerina’s Café, most of the transactions there have prices lower than $30 while there are three transactions having price higher than $55. There is also a large oscillation in this line chart. The highest price is $87.09 while the lowest is $8.29, which is next to the highest transaction. We think that these are all abnormal.

**Conclusion**

1. What does a day in the life of a typical GAStech employee look like?
2. What unusual events or patterns do we observe? For this question we need to find out the location and the person involved as well as the time when an event occurred. We also need to decide if it is necessary to pay attention to this pattern.
3. How to deal with the imperfect data?

In this section we answer the problems mentioned in the first section.

**a. What does a day in the life of a typical GAStech employee look like?**

A typical GAStech employee uses the assigned car to go somewhere near the company every day. He also goes to other places and has transactions there without using the assigned car. The number of the places visited, including the company, is less than 15. In these places Katerina’s Café often has the most times to be visited.

If he goes to Shoppers’ Delight, he is likely to have a transaction of a very high price for just one visit. The similar condition will happen when it comes to Robert and Sons and Frydos Autosupply n’ More.

When he goes to other places, the price of transactions there seems to have the total price consistent with the times these places are visited.

Those places that have the most times to be visited, like Katerina’s Café, often have transactions with various price. Some transactions have very high price while the others have low price. It seems like the employee buys different things there and some are expensive while the others are cheap.

**b. What unusual events or patterns do we observe?**

We observe that an employee sometimes goes to one place just once but has a transaction with the highest total price there.

For example, just like we mentioned in the last section, Birgitta and Ada went to Shoppers’ Delight only once but they had almost the most expensive transaction there. The time of the transactions differ from person to person, but the prices are both larger than $150. Since we have no idea what Shoppers’ Delight sell, we decide that these two behaviors are abnormal.

Besides, we also think it unusual that an employee does not use the assigned car to those places he has transactions. Maybe it is because he only park the car near the company. But we still think that this is an unusual pattern.

**c. How to deal with the imperfect data?**

We filter the data before we do the data processing work. We do adaptive sampling on the original data and delete the data of those people that do not have an assigned car. The people only visit less than 3 places are also deleted from the original data since they can hardly provide useful information.