Rails off the Rails: Visualization and Analysis of Railway Safety Incidents

Dept. of CIS - Senior Design 2012-2013

Scott Biddle
scbiddle@seas.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA

Arda Kuyumcu
kuyumcu@seas.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA

Amanda Pacheco
apacheco@seas.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA

Gorkem Yurtseven
gorkem@seas.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA

Advisor: CJ Taylor
cjtaylor@cis.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA

ABSTRACT

The Southeastern Pennsylvania Transportation Authority (SEPTA) is one of the largest metropolitan transit systems in the world, with an annual ridership of nearly 340 million [12]. A major concern with such a large transportation system is safety: each year, many safety incidents occur on SEPTA transit systems, particularly the Regional Rail Lines. Our work discusses the proposed implementation of a Geographic Information System (GIS) to be used by SEPTA. GIS help to transform collected geographic data into useful information.

1. INTRODUCTION

SEPTA is the sixth-largest commuter rail system in the country [12]. As with any large public transportation system, safety is a very high priority, particularly on regional rail. Regional rail combines the high-speed danger of subways systems, with the easy, above-ground access of bus and light rail systems into a dangerous safety risk. Regional rail trains can travel up to 90 miles per hour, and are so massive that it can take them over a mile to come to a complete stop. Sadly, once a train engineer sees someone on the tracks, there is hardly anything they can do to stop it which has caused many engineers to quit their jobs due to post-traumatic stress.

This paper starts by giving background information on SEPTA’s regional rail and geographic information systems (GIS). It then discusses related work in the field of GIS and moves forward to analyze the web application. An analysis of our web application is provided in terms of its model, the implementation techniques our group has been using, and its performance. The paper concludes with the results of the application, a conclusion of the paper and ethical challenges the web application faced.

2. BACKGROUND

2.1 SEPTA

SEPTA Regional Rail dates, as a cohesive system, to the 1950s, and comprises an extensive system of twelve lines fanning out from Center City in to Montgomery, Chester, Bucks, and Delaware Counties in Pennsylvania, New Castle County in Delaware, and Mercer County in New Jersey. SEPTA operates all of the main five mass transportation vehicles which includes buses, subways and elevated rail, commuter tail, light rail and electric trolley bus. Using these five modes of transportation SEPTA serves 3.9 million people in Southeastern Pennsylvania. SEPTA covers 5 counties while having over 300 miles of tracks which can be very hard to manage with SEPTA’s limited resources.

SEPTA safety department’s limited resources mainly refers to their transit police force. SEPTA’s police force includes 260 officers protecting all of SEPTA’s property. SEPTA controls 2,200 square miles. To put that in perspective the city of Philadelphia’s police department covers 140 square miles and 6,600 officers. SEPTA’s police force is also relatively new. SEPTA’s police force was started in 1981 which means it has only had thirty-two years of service. The city of Philadelphia is a large, busy, dangerous city, SEPTA’s police force has a lot of dangerous ground to cover and could afford some more experience. In the meantime, SEPTA’s police force needs to use their limited resources efficiently and SEPTA is looking for ways to become more efficient.

SEPTA recently joined forces with Penn Engineering and Carnegie Mellon through the University Transportation Consortium. The goal behind the consortium was bring together multiple backgrounds of engineering including computer science, electrical engineering and mechanical engineering to help improve the way the country commutes. This includes everything from safety of vehicles and the road to the analysis of traffic flow and railway deaths. SEPTA was looking for a better way to analyze railway data and looked at the connection to the consortium and Penn Engineering for help.

2.2 Increase in Rail Deaths

Railway deaths are something that often escape the public’s mind, especially considering a dangerous city like Philadelphia that has plenty of other dangers to worry about. Railway deaths have been considerably rising over recent years and safety groups are trying to spread awareness. An organization called Operation Lifesaver, started in the 1970s, has made it their mission to improve public safety regarding railways and raise awareness of their dangers. Technology
is improving at an astounding rate and Operation Lifesaver hopes to get public safety on the same level.

According to Operation Lifesaver, an organization dedicated to the awareness of railway danger, in America someone is hit by a train every 3 hours. This is a startling number and SEPTA tracks are just as dangerous and the rest of the train tracks in America. Over the past ten years, there have been 111 deaths on SEPTA’s tracks. Forty-one of those 111 were confirmed suicides which leaves the remaining seventy deaths that had an opportunity to be prevented. Rail deaths from vehicle (automobile) accidents has shown a steady decline which is promising but while this is declining there has been an increase in trespasser rail deaths in SEPTA’s area and nationally. The national total deaths rose from 411 in 2005 to 440 in 2012. SEPTA’s total rail deaths totaled at eleven in 2011 and kept rising to twelve in 2012. This is a startling difference from six total deaths in 2010 and also six total deaths in 2009.

SEPTA is looking for different opportunities of prevention but right now the only financially viable source of prevention for SEPTA is education and awareness about the dangers of railways. Right now SEPTA is using what they call safety blitzes to raise awareness which involves handing out brochures at stations that are recent hot spots for trespassing but does not have a working model of where these hot spots are. SEPTA also uses some of its education resources to address schools. When addressing railway dangers, some students replied that they walk the railways to school and they always have because the trains do not run at that time. What students and people do not know and the message SEPTA is trying to get out is that trains run at all hours of the day and from any direction. They can be on any track at any time and do not abide to a schedule. SEPTA also wants to get the news out there that trains are not what they used to be. They are a lot quieter now due to popular demand. Rail tracks are now longer and do not create the amount of noise they used to which means that trespassers may not hear the train as it is approaching in enough time to get out of the way.

Regarding the rise in rail deaths, lowering railway speeds has been discussed to reduce loss of life, even if it means losing revenue. This would be an unlikely solution to the rise in death tolls however for two main reasons. One being that cutting the train speed by half would not significantly shorten the breaking distance of trains. Typical trains range from 200 to 500 tons. Lowering the speed will make a difference but not enough to save lives. The second reason it that while it is being discussed to lower railway speeds to save lives, it is also being discussed to raise railway speeds to upwards of 220 miles per hour to decrease travel time and increase revenue.

2.3 Geographic Information Systems (GIS)

GIS is a computer system capable of capturing, storing, analyzing, and displaying geographically referenced information — information attached to a location, such as latitude and longitude or street location defined by the U.S. Geological Survey [14]. GIS is a combination of constantly improving technologies. Perhaps the first historical example of a GIS is the animal drawing — track lines and tallies thought to depict migration routes in pre-historic ages according to U.S. Geological Survey. These early records followed the two-element structure of modern geographic information systems (GIS): a graphic file linked to an attribute database. Given SEPTA’s problem of increasing rail deaths, a GIS would be a perfect tool to give their employees more information. GIS can offer SEPTA managers the ability to efficiently manage assets, maximize throughput, and especially monitor safety. When called on to deal with crime, crashes, spills, and other unusual events, railroad officials can utilize GIS for immediate access to local data such as police, fire, and rescue information, schools, hospitals, rivers, streams and other geographic information need to better manage incidents and emergencies. Currently SEPTA does not have an information system to report safety incidents, incidents meaning either trespassers on SEPTA’s tracks or people who have been hit by trains. These incidents are filed either by police or SEPTA personnel and are filed under many different categories in many different counties. A GIS could give them all of the information they need in one place.

3. RELATED WORK

Today GIS has an extensive range of use. From migration routes of polar bears to forest fires in California, from business opportunities to urban planning, geographic information systems can play a huge role in analyzing data finding efficient solutions to problems in a variety of backgrounds. Both large and small scale governments and businesses around the world are using GIS analysis to better use their data. Location intelligence, intelligence obtained by analyzing geographic data through a GIS, is an extremely valuable tool for businesses. Businesses like SEPTA are collecting copious amounts of data that can be used more efficiently. For businesses, better use of data can lead to a larger consumer base from studying current customer locations. Research in the field of GIS can lead to huge growths of companies and can help inspire business advancements. Although SEPTA is not looking to GIS to grow their consumer base, business innovation in SEPTA’s railway safety can lead to saved lives and fewer incidences by studying past incidences and using safety personnel more efficiently. The British Transport Police adopted a GIS for their police force in 2005. The British Transport Police used location intelli-
gence to pinpoint incident location which then helped them use their resources most efficiently. In this case, a GIS implementation helped the British Transport Police more efficiently serve their customers while being more cost-efficient as well. The British Transport Police have 88 police stations and over 3000 officers and using their GIS to respond quickly and accurately to emergency situations as they arise has allowed them to keep their staff at their relatively low number and respond to emergencies faster than before. SEPTA has the same problem of limited resources in regards to safety personnel and a cost-efficient GIS would be a solution. A Swedish paper published in 2005 studied railway incidences and their related statistics [11]. This study found that most suicide related train-person collisions happened during the day while most unintentional trespasser collisions happened at night. They also found that 75 percent of the suicide victims were waiting on the track before the collision. This is the type of information SEPTA is looking to possess. SEPTA has recorded this data but has not examined it. They have not analyzed when and where incidents like this happen most. With a GIS in place, SEPTA can possess these statistics and continue to analyze their data as incidences are added to their system. More importantly, a GIS implementation can help SEPTA act on their analysis. Naturally these trends will change overtime and as this happens SEPTA can study these new changes and act on it. SEPTA can rely on the mass amounts of data they possess and use it to their advantage. A GIS is an extremely powerful tool but it has its challenges. The ultimate power in a GIS lies in the data [16]. Inaccurate or not enough data can lead to an impractical, useless GIS. Data collection is problem a lot of businesses face. Common problems are conflicting data formats, reluctance to share data and lack of funding to create or buy new data. When facing these problems it can be near impossible to implement a GIS. A GIS can be a very useful tool but it often needs to be built upon. Like in Japan’s urban planning [6], GIS tools are most useful when combined with mathematical modeling and computation designed to simplify the GIS toolkit for a particular task. SEPTA is looking to adding school, business, and important building locations to the GIS and extract correlations. While facing these challenges there are many GIS resources SEPTA can call upon. ArcGIS [7], Ushihidi [15], and GIS Cloud [4] all offer geographical applications that can map incidences. These resources can be very useful but given SEPTA’s limited resources, they needed another answer. First, with SEPTA’s limited resources, which includes financial resources, SEPTA looking for a practically free GIS. All these applications have monthly fees. Given the amount of safety personnel SEPTA that would need access to a GIS, these monthly fees would be a challenge to overcome. A problem with ArcGIS for SEPTA is that ArcGIS is a desktop program. SEPTA is hoping their safety personnel could access this GIS while they are out on the job which would not be possible with a desktop.

4. SYSTEM MODEL

This system, as with any implementation project, needs to balance its requirements with its objectives. Its objectives, as described at length above, are summarized as follows:

- Visualize railway safety incident data in a flexible, user-friendly way,
- Analyze possible sources of correlation between the incident data and any possible source, and
- Display both the data and its analyses in an easy-to-understand, beautiful way.

Our requirements were as follows:

- The user interface must be easy to understand and attractive to look at,
- The results of the analysis must be understandable by the average SEPTA safety personnel,
- The application must work on many platforms.

4.1 Platform Choice

The immediate implementation strategy to address the first, most basic objective of displaying incident data in easily and flexibly, was constrained by all three requirements. It makes immediate sense therefore, to use a web-based map display. A web-based system is inherently multi-platform, and very easy to use given that the average user’s level of familiarity with the browser is greater than with a proprietary application. Given the right eye for design and the right toolkit, the interface could also be made to look any way the customer chooses it to look.

4.2 Analysis

The next objective, the analysis component, is significantly more difficult to scope and define. That objective, too, is constrained by all three main requirements, but most notably the second: no matter which metrics we use to measure correlation between sources, the results need to be easily displayed and understood in order for the analysis to make sense to this application’s target user base. Furthermore, the metrics themselves need to provide an accurate measure of the correlation and similarity between the incident data set and any potential causes. The problem, simply put, lies in both choosing the best comparison sets and displaying their output in some meaningful way.

A first, “no-brainer” metric for determining correlation between data sets is to look at the overlap between the sets and the difference in size of the datasets. Overlap between two data sets is significant as a gut-check for the very obvious reason that transportation incidents in Southeastern Pennsylvania are very likely not influenced by the locations of schools in Southern California, nor would a set of all Foursquare check-ins in Pennsylvania be of much use. In short, a simple bounding-rectangle analysis is a quick, efficient way to show if a given data set will be able to provide any interesting results. There are two components to this stage of analysis: overlap and relative size.

Overlap is a simple calculation no matter what, but its specific definition requires some thought. It is most important to determine the effect of a potential cause on the original data set, so the Overlap Test measures the percent of the minimum bounding rectangle of the incident data set which is also contained in the bounding box of the dataset being tested. The more overlap, the better, in this case. That being said, it is also very important to make sure that a high level of overlap is not caused by the cause data set being vastly larger than the incident data set, as in the example of all checkins in the United States not being particularly relevant to the problem at hand on SEPTA’s railways. Thus, the relative size between the two boxes will be
compared and factored into the overlap, such that a more similar size means “better” data. The simplest way to do so is to multiply the percent overlap by the relative size. Using this heuristic, given two datasets with identical overlap, the one closer in size to the original set will be scored as more correlative.

This simple analysis aside, more sophisticated analysis will yield more useful results. The standard statistical method to determine the similarity between (or more importantly, the difference between) two sets of data is ANOVA, an analysis of variance. ANOVA models a data set as a simpler set composed of a number of means of the data set, essentially modeling the data as a smaller set of clusters within that data set. This analysis method is well-suited to the analysis objective with a few modifications given the type of data.

The data being two-dimensional coordinates, it is important to note that the analysis must occur in both dimensions: we cannot merely compare all latitudes separately and all longitudes separately. Doing so would erase any meaningful information from the analysis. Therefore, in order to successfully run an Analysis of Variance on our data sets, the data must be clustered in some two-dimensional way. For that, we use the K-Means Clustering Algorithm.

K-Means Clustering offers a way to simplify the data set by modeling it as a (hopefully smaller) set of clusters, based on mean points of the data. The algorithm itself is NP-hard, however there are several heuristic approximations that provide a reasonable approximation. The details of implementation will be found in the implementation section of this report. What remains important in describing the model is the analysis.

Once the data is clustered to an appropriate number of mean data points, the analysis of variance can continue. Due to the data being two-dimensional, the analysis must be different from a standard ANOVA test. D.A. Griffith posited an extension of standard ANOVA practice that very nicely satisfies the need for spatial variance analysis here [8], and provides us a statistical view of how similar any two given data sets there are. The implementation of this analysis will follow in the implementation section of this paper. The end result of this analysis is either the acceptance or the rejection of the null hypothesis: the differences in the data are statistically significant.

With a simple, intuitive analysis of the data backed by a hard statistical model, the system will simply and accurately describe the degree of correlation between the two data sets. The next objective of the system then, is to display this in some meaningful way.

4.3 Information Display

Given the information the system is able to extract and the analysis it is able to provide, it must then display this information in a simple, easy-to-digest manner. There are several components to this information display, which tell a convincing story with the input data from which it is easy to grok the most important information. The first component is a map. Because the data sets this system processes and analyzes are geographic in nature, the map should be the focal point of the information display area. In addition to providing a convenient data visualization to complement the static analysis of the input data sets, the map allows users to make their own intuitive notions about the data, which is critical in providing a useful analysis tool to be used by humans. These intuitive analyses, made by experts with experience looking at that data, may continue to provide the best tools in solving the problem of railways safety incidents, and therefore should absolutely be preserved.

That being said, it is still incredibly important to provide the results of the system’s analysis. The data should be made accessible to the user in multiple formats to ensure that its meaning is to be understood. As such, the simplest and clearest ways to display them are tables and graphs. There are several design considerations to be taken into account when creating the user experience, which will be explained in the implementation system.

4.4 Summary of Model

The model consists of two primary components: the information display and the analysis supporting that display. Per client requirements, the entire system exists on the Web to satisfy the need for a compatible, scalable system. The information display contains both a map and tabular and graphical display of the analysis in order to ensure that the data is understandable at-a-glance, while still providing the capability via the map for users to draw their own conclusions based on the data displayed. The analysis itself uses two primary metrics to provide analysis: a custom, simple analysis based on very intuitive bounding-box simplification of the data to provide a coarse-grained gut check, and a finer-grained adaptation of the well-known statistical test Analysis of Variance to determine whether differences in the data are significant. This model achieves the objectives of visualization and analysis of incident and problem source data while meeting the requirements of multi-platform compatibility, effective user interface, and comprehensible analysis.

5. SYSTEM IMPLEMENTATION

5.1 Basic Layout

This system, compliant with the requirement that the platform be web-based, is written in Ruby on Rails. Ruby on Rails was chosen partly due to developer familiarity, partly due to its easy scalability and deployment on many cloud-based platforms, Heroku in particular, and partly due to its rich developer community and active library of plugins. As Fred Brooks admonishes us, when possible, software projects are always more efficient when they use off-the-shelf components instead of creating them on-demand [3], which is a principle this system attempts to follow as much as possible. Therefore, whenever possible, open-source components were integrated into the system to provide key components. They are noted where necessary.

5.2 Data Import

Data import in this system is of critical importance. Without properly-understood datasets, the system is not able to provide either analysis or visualization and therefore functionally useless. Therefore, it is important to provide an easy user experience for adding data to the system. The simplest method, reliant on the least amount of technical expertise (again, a requirement based on the system’s user base) is uploading tabular data representing various incidents. The system allows two primary formats: GeoJSON [9] for forward-compatibility, and CSV for backward-compatibility.

GeoJSON is the future of web-based geographic systems. GeoJSON is a file format that is able to describe geographic
structures, such as points, lines, polygons, and series of those in a compact, human-readable format that is also very clearly and easily parseable by a computer, and particularly by Ruby on Rails’s ActiveSupport framework [1], which provides a concise and simple tool to change JSON data in to Ruby objects and back. This file format is used by one of the main sources of data for this system, the Pennsylvania Spatial Data Access database, so developing a tool for easy conversion was practically a requirement. Currently, the tool only supports sets of points, however, the system can easily be extended to support other geographic information types.

CSV was chosen as a format because of its ubiquity and its compatibility with various and sundry spreadsheet software. The system’s average user is highly likely to be familiar with, and work daily in Excel to compile data sets. Therefore it only makes sense that this system support data input via CSV. The implementation challenge with CSV is that the data could be structured various ways with respect to the ordering of the data. The system has in place several mechanisms to determine the proper columns necessary. The first is the expectation of column headings in the first row of the data. If they exist, the system will correctly identify the latitude and longitude columns in the data and treat them appropriately. If there exists no header row, the system then tries to intuit which column is which automatically. It will read columns and try to find one with a value more than 90 or less than -90. If it can, that particular column is assumed to be longitude, since latitude values are between 90 and -90. If it can automatically find that column, it prompts the user to confirm the correct alignment of the columns. Failing that, the system will prompt the user to select the correct data labels for latitude and longitude.

As this tool is a static analysis tool, the imported data is stored in memory and not saved to a database. For the user’s convenience, the user may optionally elect to save incident data to the application's database for ease of use in repeated analyses.

5.3 Analysis

Once the data is properly imported, the system can begin its analysis. The first step, as discussed above in the System Model, is to determine bounding box overlap and relative size between the several data sets. The calculation of bounding boxes is not particularly difficult; the bounding box is defined as having an upper edge equal to latitude of the northernmost point, a lower edge equal to the latitude of the southernmost point, a right edge equal to the longitude of the easternmost point, and a left edge equal to the longitude of the westernmost point. These determinations can be made very simply and in linear time by a simple iteration over each data set. The comparison calculations are equally simple, requiring only a few lines of code to determine relative size and percentage of overlap. These two metrics are, as discussed above, multiplied to yield an overlap score.

The Analysis of Variance testing is significantly more complex. There are two primary components to the analysis: the k-means clustering and the variance analysis based on that clustering. In order to complete the clustering, the analysis makes use of Red Davis’s K-Means Clustering framework for Ruby [5]. The framework takes as input a series of pairs (in the case of this system, latitude and longitude pairs), and the number of centroids, or clusters to end up with, and yields a series of converged mean cluster points. The number of clusters is given as a parameter to the user to decide. In order to make the difference apparent to the user, the number of clusters in the analysis is shown as a slider, where the user must choose between “faster” analysis (3 clusters) and “more thorough” analysis (10 clusters).

This clustering algorithm is repeated for each data set, and then the analysis of variance is conducted. The variance analysis proceeds much in the manner of Griffith’s paper, cited above, implemented in Ruby. The end result of the algorithm is, as stated above, an acceptance or rejection of the null hypothesis that the differences in two data sets are statistically significant. That acceptance or rejection is remembered for the data presentation component of the system.

5.4 Data Presentation

There are three primary components to the data presentation component of this system. The first is the map component. In order to effectively display the map in a web based system, Rails off the Rails uses Leaflet.js [2]. Leaflet is a lightweight, performant, Javascript-based map framework with map data provided by OpenStreetMap. There were several options in choosing a mapping framework: Google Maps, Bing Maps, FacilMap, and others. This system uses Leaflet for many reasons, first and foremost is its flexibility. The framework allows for tile layers, markers, and lines, which are exactly what we need for this analysis tool. Furthermore, the framework is highly optimized and efficient in its display of the information, and allows for the quick addition and removal of various data layers. This feature is a key determining factor because the ability to add and remove data sets from the map both quickly and efficiently is a critical component of the system's ability to allow its users to perform their own intuitive analysis on the data.

Another determining factor in the choice of Leaflet.js is its use of OpenStreetMap to provide map data. OpenStreetMap is a free and open-source, collaborative world map project which offers accurate, up-to-date maps free of charge. OpenStreetMap provides all of the data this system needs to be effective while also conforming to the customer’s preference for open source, free-as-in-beer and free-as-in-speech components as well as the development team’s budget of zero dollars.

A crucial element in improving the map’s readability is the ability to distinguish points from various data sets. There are two major implementation details that are responsible for providing a consistent and easy-to-understand user interface in this respect. The first is Leaflet’s ability to customize markers for various points. Markers can be styled in myriad ways, so each data set gets points of its own color, randomly assigned from a small selection of colors selected to contrast with each other so that at a high level, the several data sets displayed at any given time on the map can be easily distinguished. The second element is a plugin provided by the Leaflet Community: Dave Leaver’s excellent Leaflet-markercluster plugin [10]. The clustering allows for a very very intuitive grouping of marker data by region based on the current zoom level of the map. However, clustering every single data layer (the incident data as well as all potentially-correlated data) would obviate the need for separate coloration and make intuitive analysis very difficult. Keeping this in mind, the system only clusters the incident
data, allowing for both intuitive analysis of that particular data on its own, and also analysis based on the incident data layered with any potentially-correlated data.

Because the map is the focal point of the user interface, the experience was designed around the map. The map is large, taking up three-fourths of the width of the user interface. On either side are controls to alter the experience of the map. A simple multi-select control, common and easily understood to users of all technical abilities, to add and remove layers of data from the map. The layers are named and a color-based legend is provided to provide simple and easily-understood explanations of the map data.

Beneath the map lie the tabular results of the system’s analysis. The tables, while useful in and of themselves, are made simpler for non-technical users by the simple and subtle use of color. For the results of the bounding-box Overlap Test, the data set with the best overlap score is highlighted in green, and the worst score is highlighted red for at-a-glance understanding of the data. The results of the more sophisticated statistical analysis are presented with a simple Significant/Insignificant output, as well as a corresponding green or red highlight.

The tabular data can also be displayed graphically. The results of the statistical analysis are little more than a slightly different representation of the tabular data, as their output is essentially a boolean variable, but the results of the bounding box Overlap Test are sometimes better represented graphically. In order to create graphs, the system uses a Javascript-based framework called tufte-graph [13]. Tufte-graph allows tabular data to be represented graphically with little additional work, so it is a natural fit for this application. Furthermore, it is also free and open-source which fits the needs of this system.

5.5 Summary

The system consists of a data import, analysis and presentation component. The import component uses a straightforward interface and user-confirmed intuition in order to ensure that the correct data is being analyzed. The analysis component uses both simple and complex statistical models to determine correlation between the several potentially-correlated data sets and the incident data. The presentation component focuses the user experience on the map, but also provides simple explanation of the complicated analysis to ensure that users of all backgrounds and technical experience are able to understand the data presented by the system. These design decisions satisfy the customer’s requirements while simultaneously meeting the objectives of the system.

6. RESULTS

When the client SEPTA approached the project team through Penn’s collaboration with the University Transportation Center, the safety personnel had various ideas in mind. Some of the ideas were that there existed a potential correlation between the number of incidents in a given area and the number of schools and office buildings in that area. If this assumption held, SEPTA would easily be able to identify schools and buildings of relevance, and focus their educational efforts towards people in there. The system has been built upon this initial guidance, and the instinct was to place incident data on the same layer as schools in Pennsylvania. The geographical data clearing house called Pennsylvania Spatial Data Access (PASDA) came in very handy for this task. Through using PASDA, access was possible not only for data regarding schools in the state, but also for demographic information and population of the region.

The findings however did not accurately align with the direction SEPTA provided initially to the project. Instead, the findings turned out to be relatively unusual and close to what one would expect; the most incidents occur in the SEPTA stations that are the busiest and most frequently crowded. In the past decade, the largest number of incidents happened in the Center City region of Philadelphia. The City Hall Subway Station is one major station of interest, as our findings showed that about 25% of all incidents were reported there. A potential proactive measure that could be taken in light of this situation is creating educational pamphlets and brochures, and distributing them at busy stations such as City Hall. Awareness could also be raised by using some of the advertisement space in wagons of the subways that run on busy lines such as Market-Frankford Line.

Areas outside of Philadelphia were not found to be of high interest to SEPTA’s awareness efforts in general. To further this analysis, the team tried to focus on a busy station outside Center City. Paoli Station was a good fit for this description, as it services not only SEPTA subways but also Amtrak for the suburbs. Yet hardly ever any incidents happened in Paoli Station, mainly because it is still a smaller station compared to 30th Street Station which also handles a lot of traffic from Amtrak, NJ Transit and SEPTA. A possible explanation for this phenomenon is certainly the demographics of Paoli; one could argue that the suburban culture is a reason why there are less incidents in Paoli Station. Yet, it is noteworthy that the demographics of a region was not found to be correlated with the number of incidents in that region. This is a relatively more surprising finding, as one is likely to suspect otherwise. However, a busy station such as City Hall hosts traffic from the urban population along with a large number of commuters from the suburbs everyday. It is important to mention that at this point, it is up to the safety personnel to draw more educated conclusions and analyze the information more thoroughly.

In terms of speed and flexibility, the application performed well. Several factors play into this result. The first one was the decision to use leaflet.js and OpenStreetMap as the mapping framework over Google Maps. The combination of leaflet and OpenStreetMap provided sufficient functionality tailored to our needs and kept the application lightweight at the same time. Even though Google Maps offered plenty of capabilities, the complex nature of its API and the inclusion of many functions not required of our application made it hard for the team to work with it. Another contributing factor to performance was the size of our data set. Thankfully, no more than 4000 incidents have been reported in the last decade which made for a manageable size speaking in terms of computer science. The application is able to host this data statically, which in turn provides faster access and analysis on it. A potential challenge for SEPTA moving forward could be the maintenance of such performance when the integration with its database for real time updates is complete.

7. FUTURE WORK

The system built for this project is meant to be a generic and scalable one. Ideally, a person who previously has not
interacted with the system should be able to upload data and draw correlations from the output. The current data parser is built in a specific way tailored to the data that SEPTA provided. The incident information included in the data is the date and time of the incident, the coordinates and address of the information, and whether there was an arrest made or not. In the future, the parser component may be transformed into a more generic one in order to provide general capability of converting data into JSON format. Moreover, the system is built to provide information for incidents in the Greater Philadelphia Area. This however is not a limitation, as the mapping frameworks used to build this system have full coverage nationwide.

On a greater scale, there are further improvements SEPTA can possibly make to the project. When SEPTA takes over the application, the technical personnel should certainly look to integrate the system with their database as a first step. By doing so, safety personnel at SEPTA will be able to see real time incident updates. If a mechanism to accomplish simple reporting is not in place, SEPTA may certainly develop a simple framework to digitize incident reporting. Another approach that could be proposed for the implementation of incident reporting might be input from the general public on various media. Through using different channels such as social media and hotlines, residents may be encouraged to provide input and inform the authorities. However, there are downsides and ramifications to the use of this approach. Technical challenges to this approach might include the necessity of consolidating various input to a single standard format the database at SEPTA follows. Additionally, the legitimacy and reliability of the information provided by the public might not always be very high, and SEPTA would have to be aware of this.

The system might be further automated with the inclusion of machine learning techniques to predict and prevent railroad incidents. Even though the client did not specify such automation as a requirement for the application, the usage of simple techniques such as Naïve Bayes and perceptrons may provide better insights to the safety personnel by categorizing incidents. SEPTA may also expand the usage of the application to include information regarding where people check-in to on social media. Through plotting the relevant coordinates on a mapping framework, visualization of the data has been implemented. The visualization aspect that personnel can simply focus at the map and focus on proximity to the location of frequent incidents by visualizing them. By doing so, SEPTA will be able to draw correlations (but not necessarily causations) between incidents and other factors. This is important for the safety personnel at SEPTA specifically because the region SEPTA supervises is simply too large for the number of safety employees there. By concentrating their efforts at areas that are most pertinent to incidents, SEPTA will run focused awareness campaigns and minimize the frequency with which accidents occur in the most efficient manner.

8. ETHICS

Once SEPTA takes over the development of this project, they have an important decision to make. It is up to SEPTA to decide whether they want to make the incident data publicly available or not through this web application. While making this application accessible by the public might help in terms of coming up with better results, it might also lead to ethical challenges such as unique identification and intentional corruption of data. If SEPTA does indeed build a system through which residents can report incidents, a big challenge arises. SEPTA has to be able to trust the residents of Pennsylvania that such a system won’t be misused.

Through the use of open source technologies, the application inherently adopts licenses provided for the use of developers and the general public. The implication of this is that our team and SEPTA may legally take ownership of the project, including the user interface and the data analysis tools built for this project. However, it should be made clear that this ownership is only a symbolic one; in order to contribute to the developer community, it is our duty to make the project as publicly available as SEPTA’s privacy constraints allow us to. Moving forward, SEPTA may certainly expand upon and further improve this project as they prefer to, without necessarily releasing the incident data that belongs to them.

9. CONCLUSION

The Federal Railroad Administration Office of Safety Analysis reports 254 total deaths between the years of 2004 and 2012 on railroads in the state of Pennsylvania. We aim to help SEPTA to concentrate their efforts in implementing preventative measures against railroad safety incidents. The main goal is to identify areas of most interest in terms of proximity to the location of frequent incidents by visualizing them. By doing so, SEPTA will be able to draw correlations (but not necessarily causations) between incidents and other factors. This is important for the safety personnel at SEPTA specifically because the region SEPTA supervises is simply too large for the number of safety employees there. By concentrating their efforts at areas that are most pertinent to incidents, SEPTA will run focused awareness campaigns and minimize the frequency with which accidents occur in the most efficient manner.

As a result of this project, the goal of building a Geographical Information System was achieved. Every system of such nature has a common goal in mind, which is to transform data into usable information. The web application successfully integrates data from several different sources, including incident locations, school locations and locations people check-in to on social media. Through plotting the relevant coordinates on a mapping framework, visualization of the data has been implemented. The visualization aspect is especially important for the intended audience, which are non-technical SEPTA safety personnel. The convenience is that personnel can simply focus at the map and focus on particular areas of interest as opposed to going through approximately 4000 data points on a spreadsheet. By abstracting away the technical details behind a clean and simple user interface, the application welcomes users from any range of prior technical background.
It is known that there are commercial Geographical Information Systems out there that offer similar functionality or even better. However, this application has numerous advantages over others as well. The system has been built from ground up using completely open source tools and frameworks. Many advanced GIS products are available at a high cost and require maintenance, whereas the system built offers the bread and butter functionality for free. Additionally, the most prominent examples of GIS products in the professional field come as desktop applications. The underlying implication of this is environment dependencies and lack of portability from one operating system to another. Our application is developed for the web, and is accessible from anywhere one can find an Internet browser. The platform independence brings high mobility and portability to the application, which is an advantage over commercial products. Finally, the aforementioned argument regarding ease of use for individuals with non-technical backgrounds applies here as well. Commercial systems such as ArcGIS by Esri [7] introduce the necessity of professional support teams and hotlines at all times, largely because of the complex user experience they provide. Our system maintains a nice balance between simplicity and usefulness in analysis and conveying information.

10. REFERENCES

http://api.rubyonrails.org/classes/ActiveSupport/JSON.html.


http://www.geojson.org/geojson-spec.html.

https://github.com/Leaflet/Leaflet.markercluster.


http://xaviershay.github.io/tufte-graph/.

