Final Report

IU Foundation | United States Freight Rail Vulnerabilities

Project Sponsor:
Jason Gumaer
Kyle Stirling
Joy Zayatz

Project Members:
Allen Ho (allho@iu.edu)
Erik Gonzalez (eriagonz@iu.edu)
Jonathan Hansen (jnthnhansen@gmail.com)
Samhitha Nuka (snuka@iu.edu)
Shawn Strasser (sstrass@iu.edu)
Yashwanth Vijayaragavan (yavijay@iu.edu)

1. Overview

The United States rail system plays a critical role in transporting freight across the nation, moving approximately “1.6 billion tons across nearly 140,000 miles of track” (AAR, 2024) throughout the course of a normal year. The materials moved via this rail network vary widely, including categories such as live animals, fertilizers, chemicals, vehicles, and more. In the case of an incident - such as the East Palestine, Ohio derailment in 2023 - communities surrounding rail networks can face various economic, environmental, and human impacts.

This project aims to help communities better understand the potential risk they face from such a rail accident The solution was provided in the form of a risk index using a one through ten scale. This risk index currently consists of three primary components - the likelihood of a rail accident occurring in a given track segment over a certain period of time, the social vulnerability of the surrounding population, and the expected type of commodity volume for freight in that given location - that are combined into a single number. A machine learning model was developed to predict the likelihood of an accident utilizing historical rail accident data made available by the Federal Railroad Administration, and rail preemption events at signalized intersections provided by Indiana Department of Transportation. The Social Vulnerability Index provided by the CDC was leveraged to determine the risk to communities surrounding rail. Commodities being trafficked on rail for a specific location were estimated by combining data from the Freight Analysis Framework (maintained by the Bureau of Transportation statistics) with US Census data. This project focuses on Indiana, but the underlying datasets are provided at the national level, allowing for scalability. The risk index was developed in a way where future work can improve it by either adding in more components, or further refining those which are already provided.

2. Purpose

The East Palestine, Ohio rail incident which occurred on February 3, 2023 caused significant harm to the community and environment surrounding the site of the derailment. This train was transporting hazardous materials such as “vinyl chloride and butyl acrylate chemicals … chemicals [which] are used in manufacturing plastics and resins” (Idem, 2023), causing particular concern when these rail cars ruptured. A timeline of the incident notes that it led to a four day long mandatory evacuation of homes within a one mile radius of the incident, and to the water in the surrounding area not being declared safe to drink until February 15th, twelve days after the derailment (Schnoke et al., 2023). The health impacts of this derailment are still to be determined, with the National Institute of Environmental Health Sciences funding six studies to identify and track impacts following the accident (*East Palestine (Ohio) train derailment research response*).

While the East Palestine incident was well publicized, many more derailments happen throughout the US rail network in a given year. Given the breadth of this network and the

relative obscurity surrounding most derailments, it's unlikely that most individuals residing near a rail network have spent much time considering the risk they may face from a derailment. This can be further extended to the potential risks associated with the materials being transported near these communities via rail.

This project aims to help communities/individuals better understand the potential risk they face from such a rail incident over the course of a 10 year period. The goal is to provide this information via a geospatial visualization, while also providing end users with the context to properly interpret the provided metric.

3. Methodology

This semester's team aimed to build upon the work of past cohorts, which had identified relevant accident data sources and created interactive maps displaying historical accident data. This work was furthered as this semester's project team aimed to determine an appropriate method or metric for communicating risk, and develop a framework which future cohorts can expand upon. For the purpose of this analysis, incidents do not include accidents at rail crossings - for example, when a train collides with an automobile. The methodology to do so consisted of:

1. Determining the appropriateness of a Risk Index
2. Exploring available datasets for feature development
3. Risk index component development and aggregation
4. Creating a visualization which communicates this information
5. Establishing a framework for future semesters to add onto the Risk Index
    a. Existing components can be improved
    b. New components may be added and incorporated
    c. The scope of the project may be expanded beyond Indiana

4.1 Determining the appropriateness of a risk index

When determining the most appropriate method for addressing the problem posed by the project, the team identified multiple potential problems that could be worth addressing. One such problem was the likelihood of an accident occurring. Another was the likelihood of a train containing hazardous material. Finally, quantifying the risk to the surrounding population was identified as a key aspect of answering the problem posed by the project. As each of these elements had relevance to the project problem statement, it was determined the best path forward would require identifying a method to combine them. A member of the project was able to provide the team with research regarding how to develop and implement a risk index in the context of identifying workplace risks, and this concept was identified as the best path forward. A risk index allows the team to develop each of the relevant features and thoughtfully combine them into one useful metric.

4.2 Exploring available datasets for feature development

The decision around which features to use was informed by the data that was available. Accident data was readily available on a national scale, provided by the Federal Railroad Administration (basecamp link). This site has accident information available for rail crossings - intersections where tracks cross the road - and for non-crossing accidents.

Rail preemption events at signalized intersections were provided by the Indiana Department of Transportation (basecamp link). Each time a train passes through a signalized intersection, it sends a preemption request to the traffic signal controller, which logs the event. This data was used to estimate train volumes for each track segment, which in turn was used as part of a model to estimate the probability of a train accident on each track segment.

When determining the risk to the surrounding population, research demonstrated the Social Vulnerability Index created by the CDC to be a suitable metric (basecamp link). Suitability is expounded upon in the Impact/Outcomes section.

Attempting to identify commodities traveling along rail proved to be difficult, as there is not robust publicly available data documenting the flow of commodities along rail lines. The team leveraged data from the Association of American Railroads showing the flow of commodities to and from states (basecamp link). This was useful at a state level, but a slightly more granular dataset, the Freight Analysis Framework provided by the Bureau of Transportation Statistics (public link,) contained data showing the flows of commodity volume between metropolitan areas and states, offering a more granular view of the data. Due to the dataset size, a python script was used to reduce the dataset to only relevant Indiana volume. The Freight Analysis Framework data was preferenced when available. This commodity data was combined with US Census employment data, to further increase the level of detail regarding where commodities in the state are likely flowing.

4.3 Risk Index Component Development and Aggregation

The various datasets were leveraged to develop the distinct features of the risk index, and the process for doing so is described in greater detail within the Results section of the report.

4.4 Creating a visualization which communicates this information

Once the risk index values were created and combined for each geographic location, they were incorporated into an interactive visualization. A sample of the visualization can be found in the Impact/Outcomes section.

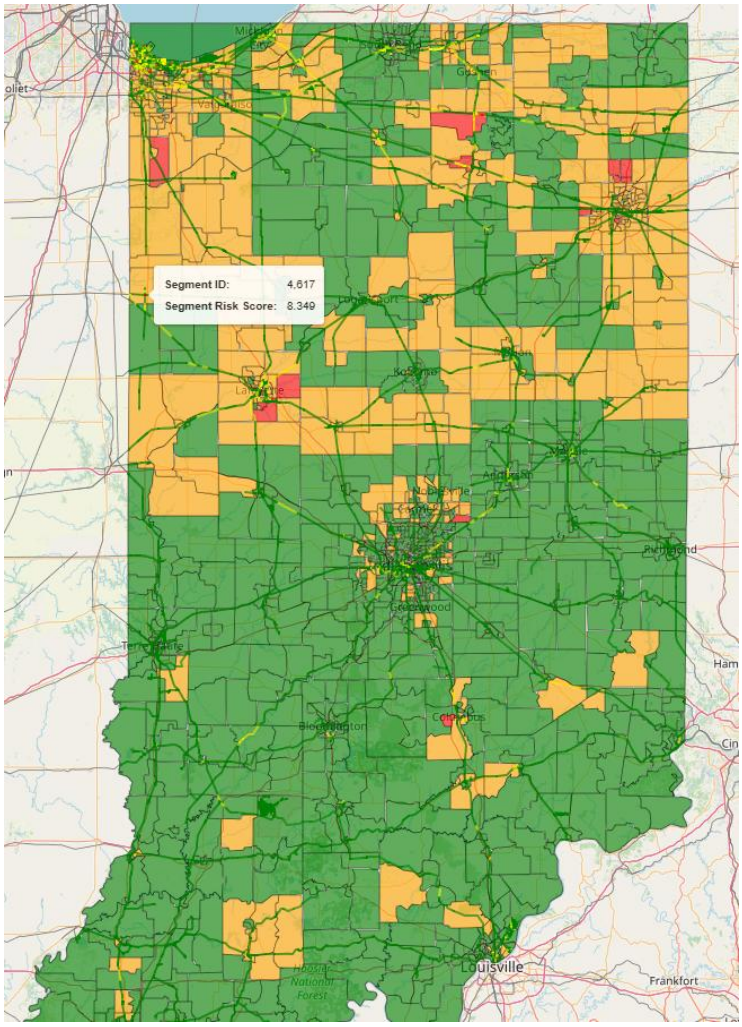
4.5 Establishing framework for future semesters to supplement Risk Index

Given the short nature of the project when compared to the scope of the problem, the team prioritized developing the solution in a way that future cohorts could pick up and contribute to. Details of this approach are available in the Recommendations for Future Work section.

5. Impact/Outcomes

The deliverable is provided in the form of a risk index, accessible through an interactive map, which individuals or communities can use to look up and determine the potential risks they may face from a rail incident. Knowing their relevant risk index, individuals and communities can then determine if there are possible steps they should take to reduce their risk, if needed.

The risk index is provided on a one through ten scale for ease of interpretation. An example of the visualization can be seen below, and the interactive map can be found here.



6. Results

The practical deliverables for this project (aside from the visualization) included the development of an accident prediction model, integration of the Social Vulnerability Index, development of a method for estimating commodity volume, and combining these features into a single risk index.

6.1 Estimating the Likelihood of Rail Accidents

To estimate the likelihood of a rail accident we used two datasets. The first was the volumes dataset and the second was from the Indiana Accidents Dataset (linked here).

Volume Dataset

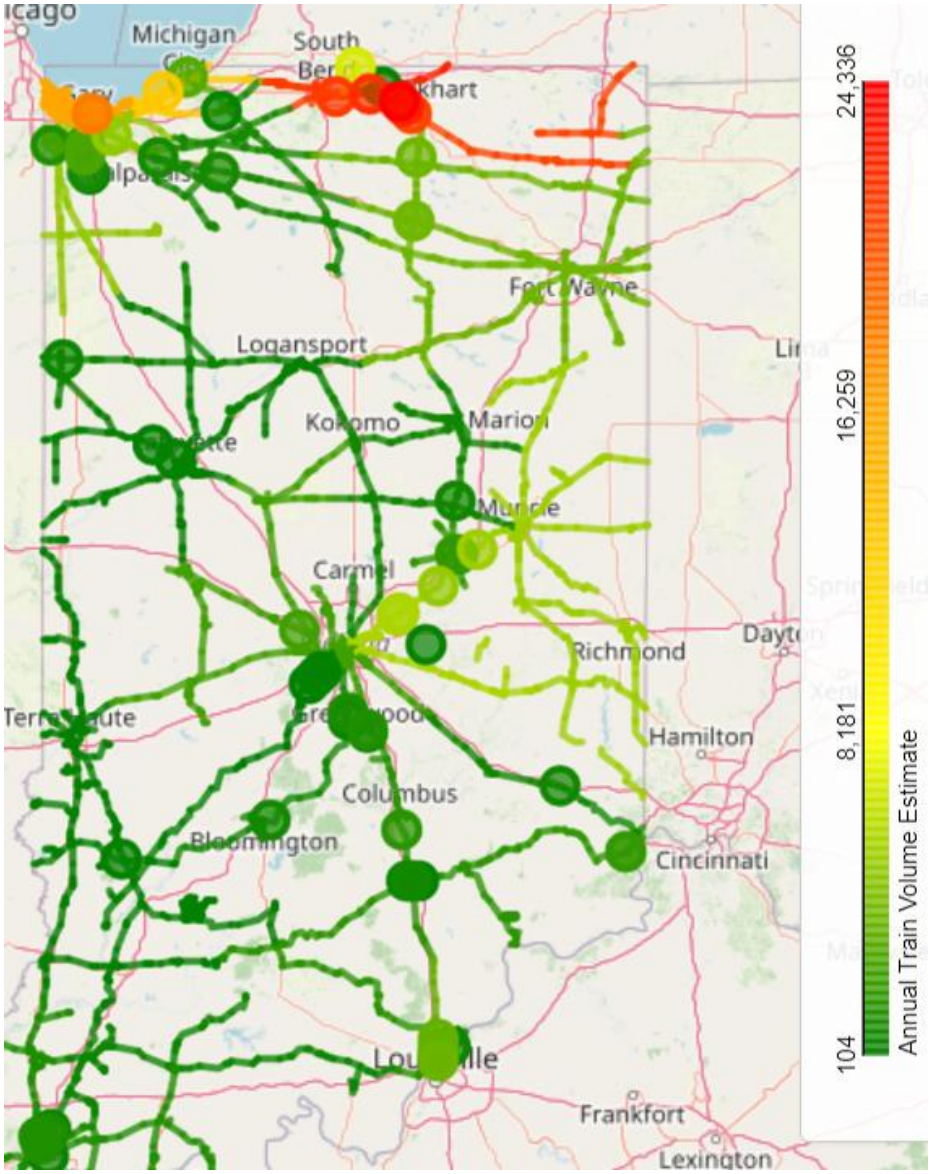
The volume dataset estimates the annual number of trains traveling over each track segment. These estimates were derived using preemption event logs from the Indiana Department of Transportation (InDOT), which record when trains activate traffic signal preemption at signalized intersections.

The preemption event logs (railquery_2024-10-22.csv) include industry-standard event codes that indicate the type of event. We focused on event code 106 (Preemption Begin Track Clearance) to isolate train preemption events and exclude other types such as emergency vehicle preemptions.

Steps to process this data:

- **Event Filtering**: We filtered the dataset to include only event code 106, ensuring accurate representation of train crossings.
- **Data Transformation**: Using the SignalTimer App (available at signaltimer.com), we transformed the event logs to calculate the frequency and duration of train crossings at each intersection.
- **Aggregation**: The filtered events were aggregated to determine the average daily train counts at each signalized intersection.
- **Mapping to Track Segments**: The process of assigning train volumes to track segments used an iterative spatial joining approach. First, volumes from the closest traffic signal points were assigned to track segments within a 0.001-degree threshold distance. For remaining unassigned segments, the process was repeated with gradually increasing distance thresholds until all track segments received a volume estimate. This iterative method ensured more accurate volume assignments for tracks near signal points while still providing reasonable estimates for more distant segments. The final output was exported as a comprehensive dataset containing track segment geometries with their corresponding annual volume estimates.

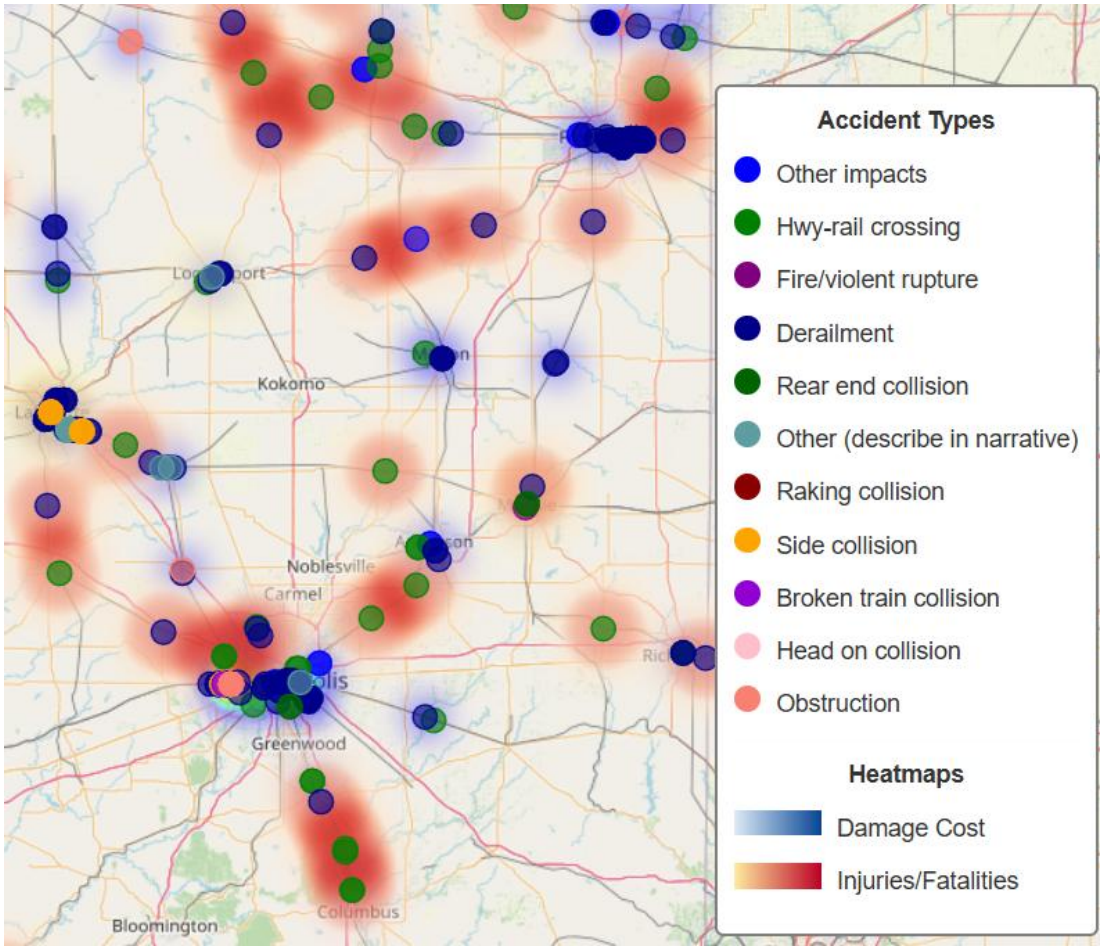
The code and data used for this analysis are available in the project's GitHub repository: [dsci-d592/US_Freight_Rail: DSIP Project]



Annual Train Volume Estimate (full notebook in GitHub)

Accidents Dataset

The accidents dataset comes from the United States Department of Transportation, and includes detailed descriptions of rail accidents since 2011. It includes details such as the accident type, any injuries or fatalities, and associated damage costs. Below is an example map of these accidents.



Accidents (from Accidents_Visualization.ipynb in GitHub)

The first steps in refining the Indiana Accidents dataset beyond basic cleaning such has dropping duplicates and dealing with null values were to convert the datetime to a usable formatting, create the geometries used for locating the track segments using shapely and then converting the dataframe to a GeoDataFrame. The features used from this dataset were Accident Number, Year, Month, Day, and geometry. The accident geometries were buffered by 50 meters to allow for slightly incorrect data in the GPS coordinates locating track segments.

The train volumes dataset also converted the geometry column to geometric object using shapely and transformed with GeoDataFrame for compatibility with the accidents dataset.

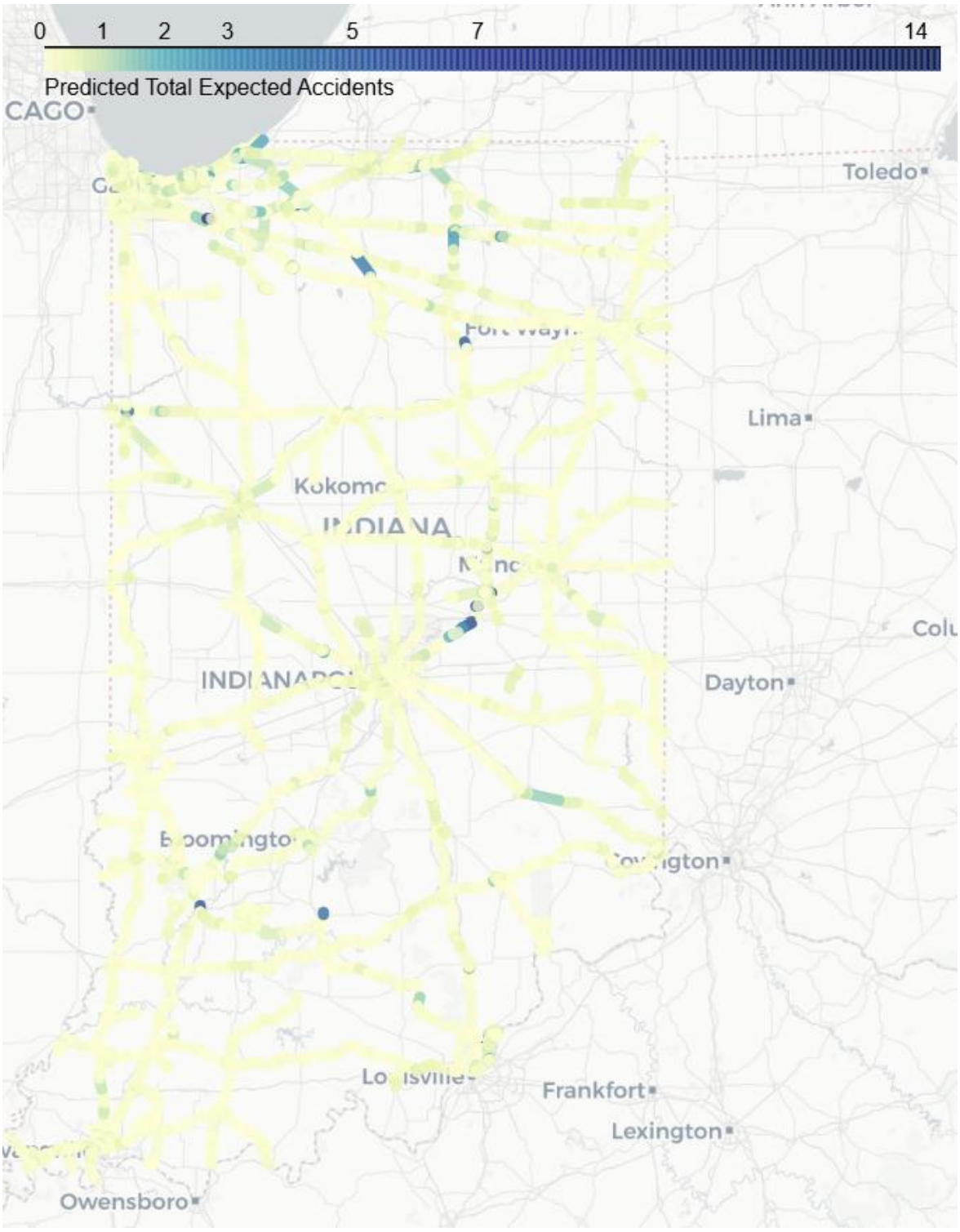
A spatial join was used to combine the accident and volume dataset and assign accidents to the appropriate track segments. Since many of the track segments had 0 total accidents since 2011, and because there will always be some probability that an accident would happen steps were taken to provide a baseline. In order to assure that none of our track segments had a zero value, the total accidents in all of Indiana were divided by the total volume in all of indiana. This number, which was small ~ 0.017 was overlaid onto each track segment. Then the actual accidents on each track were combined with the baseline to get our total expected accidents for each track segment. Before implementing the machine learning model a log transformation was

applied to the annual estimates and miles, polynomial features were added to help with non-linear features, the data was standardized, and outliers were removed.

In order to find a model that worked well 5 different models were tried. The first round of running the models received very little predictive power, however after adding some of the transformation techniques mentioned above, incorporating additional features (number of tracks, rail yard or not, passenger track or not), and removing outliers this was the output:

```
Model Performance Summary:
Poisson Regression RMSE: 1.19, R-squared: -0.07
Enhanced Poisson Regression RMSE: 1.10, R-squared: 0.10
Random Forest RMSE: 0.95, R-squared: 0.40
GBM RMSE: 0.96, R-squared: 0.39
LightGBM RMSE: 0.93, R-squared: 0.43
Neural Network RMSE: 1.00, R-squared: 0.33
```

While the LightGBM was the best performing model we ended up selecting the Random Forest Regressor due to simplicity of implementation. The accidents dataset was built from data collected from 2011-2024. Taking this data we came up with a daily prediction for each track and then expanded that prediction to annually.
Here is the map created outlining the track segments more prone to accidents based on the Random Forest Regression Model:



Additional features could help make this prediction even more powerful. For example accounting for speed by filter for rail yards, curves, weather patterns, or other variables that could cause more accidents.

6.2 SVI data ingestion and integration

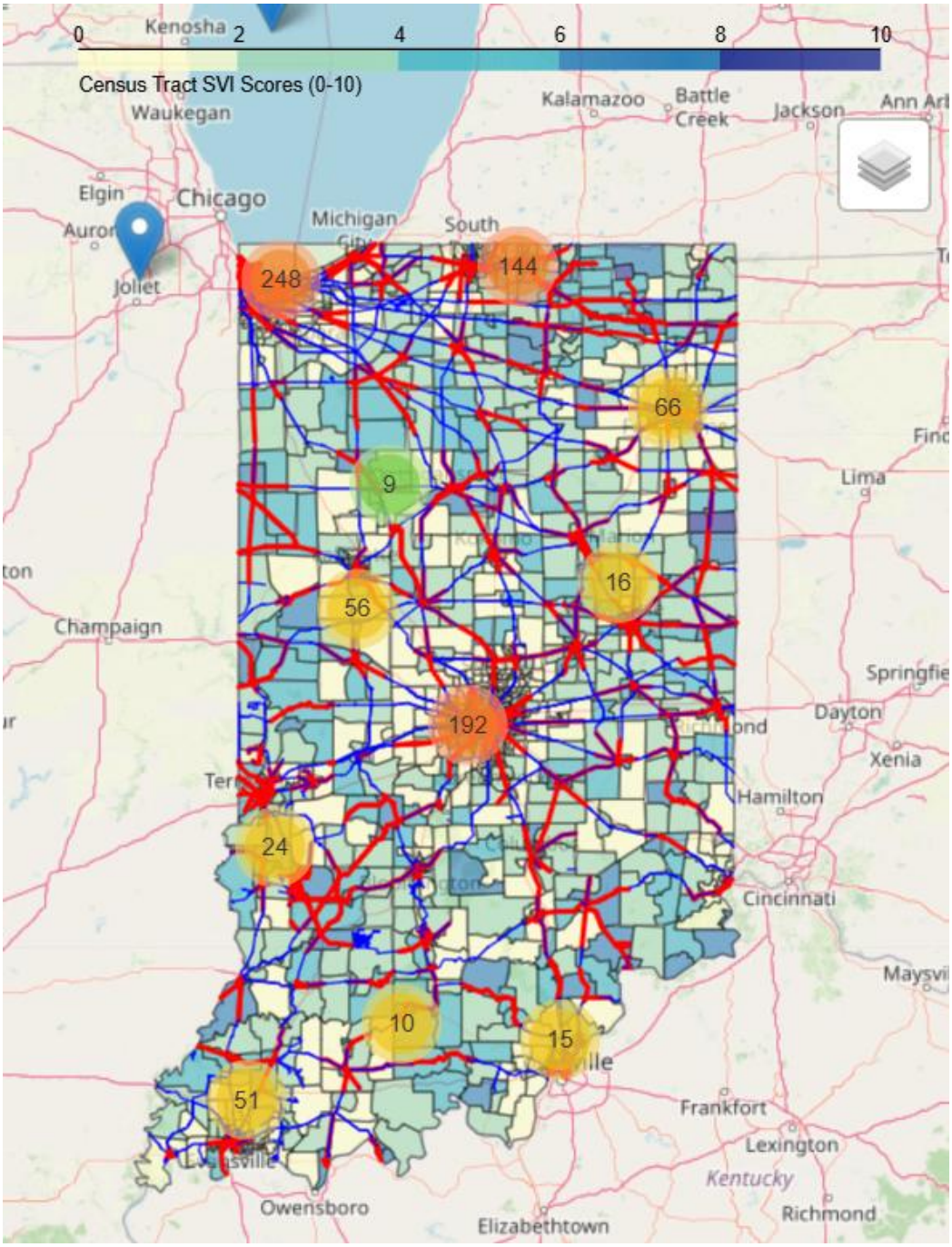
The Social Vulnerability Index (SVI), developed by the Centers for Disease Control and Prevention (CDC), is a critical tool for assessing the vulnerability of communities. By evaluating a community's ability to prepare for, respond to, and recover from stresses like natural disasters, economic disruptions, and public health crises, the SVI provides actionable insights. It aggregates data from key social factors such as socioeconomic status, household composition, race/ethnicity, and housing characteristics to generate a comprehensive score for each census tract. This score, ranging from 0 to 1, highlights areas most likely to experience adverse outcomes during a crisis, allowing for targeted interventions. The SVI's credibility is underscored by its widespread use in disaster management, urban planning, and public health, making it a robust foundation for our analysis.

The SVI encompasses four major themes, synthesizing data from the U.S. Census to identify factors that heighten community vulnerability:

- Socioeconomic Status: Includes poverty, unemployment, income, and education levels.
- Household Composition: Accounts for age, disability, and single-parent households.
- Race/Ethnicity and Language: Highlights minority status and language barriers.
- Housing and Transportation: Evaluates housing type, crowding, and access to transportation.

In our project, the SVI played a pivotal role in identifying and addressing vulnerabilities within communities. The index allowed us to pinpoint regions that are socially and economically disadvantaged, helping to assess the risks posed by external factors like train derailments and disasters. By identifying these high-vulnerability areas, we ensured our efforts were focused where they were most needed. Additionally, the SVI informed decision-making around resource allocation, guiding stakeholders to prioritize aid and mitigation strategies for at-risk populations, thereby fostering equitable responses to potential crises. Integrating the SVI also added a crucial layer of social context to our risk models, enabling us to analyze how vulnerabilities interact with physical risks and to develop more comprehensive and actionable assessments.

The SVI has demonstrated its utility in real-world applications such as disaster management, urban planning, and public health. For instance, it has been used to prioritize emergency responses during hurricanes, guide infrastructure development in under-resourced areas, and ensure equitable vaccine distribution during public health emergencies. These examples highlight the index's versatility and relevance, reinforcing its importance in our project. By leveraging the SVI, our work aligns with best practices, ensuring that our recommendations are both data-driven and socially conscious. The integration of the SVI into our analysis allowed us to address not only physical risks but also the societal inequities that exacerbate them, thereby maximizing the impact and fairness of our interventions.



6.3 Commodity Data Risk Estimation

Commodity risk taken from three methods of estimation. Census data (linked here and here) in combination with commodity data taken from the Association of American Railroad (linked here) was used to estimate commodity data based on geographic census location of indsutry data.

Industry location data was estimated by used as an assumption that employes lived near their work place.

In order to further refine the state level volume estimates provided by the AAR data, the Freight Analytics Framework (FAF) data containing city level data was ingested. This data was used for two purposes. The first purpose involved identifying volume that either originated or terminated in a city within Indiana, to provide an updated distribution to supplement the AAR data. In addition, it was recognized that some portion of the rail volume traveling through Indiana neither originates nor terminates in the state. To estimate the distribution of this volume, logic was developed to identify what national volume is likely to pass through Indiana by drawing straight lines between origin and destination cities in the dataset, and checking to see if those lines intersect any Indiana counties. If they do, this volume was attributed to the cities associated with those counties. Code to refine the FAF data can be found here, logic to identify relevant passthrough volume can be found here, and logic to identify the relevant terminated volume is available here.

Common commodity categories were created across the datasets and each assigned a risk score between one and ten. These risk scores across the various commodity types (pass through volume and census distributed terminated volume) are then combined. The weighting for the types of commodity volume can be changed, but are currently set at equal weight. This combined weighted score is also provided as a value between one and ten.

6.4 Combining Components into a Risk Index

Outputs from the above components were converted into a consistent format and combined into a single risk index here, using the formula specified below:

$$Risk\ Index\ =\ w_{AL} * Accident\ Likelihood\ +\ w_{cmd} * Commodity\ Risk\ + w_{svi} * SVI\ Value$$

The results yield a risk index on a one through ten scale. Each individual component is provided on a one through ten scale. The weights sum up to a total value of one. This ensures that the final risk index value will also be provided on a one through 10 scale.

$$w_{AL}\ =\ Set\ to\ 0.5$$
$$w_{cmd}\ =\ Set\ to\ 0.25$$
$$w_{SVI}\ =\ Set\ to\ 0.25$$

4. Conclusion

In the current environment, there is not currently significant visibility regarding the prevalence of rail accidents, nor a clear understanding of the freight aboard the trains and how an accident can potentially impact a surrounding community. To help solve for this, the team has built on the work from past semesters to develop a risk index. This index provides individuals and

communities with a better understanding of the risks they may encounter from a rail incident within their geographic location. This risk index was designed to leverage a simple framework - one that currently consists of three components - such that existing components can be enhanced or new ones can be added by future semesters. The current output includes a map with accessible risk indices by geographic location. Future semesters should review and refine the methodologies behind existing components (rail accident prediction, commodity estimation, and SVI) or consider adding new elements to the risk index. Extending this work will improve the robustness of the risk index.

5. Recommendations for Future Work

The work completed this semester was intended to also create a framework that future cohorts could build upon. Leveraging a risk index allows for future cohorts to tackle distinct aspects of the problem while combining them with the work that has been done before. All summarized recommendations have been listed below:

1. Develop and incorporate new features into the risk index
    a. Population Density, likelihood of other natural disasters, proximity to waterways
2. Enhance the accident prediction ML model
    a. Can features such as track geometry, max speed, etc… be used to improve the prediction capabilities of this component
3. Improve the commodity estimation methodology
    a. The work this semester leveraged assumptions to aid in developing this feature; these assumptions can be validated or enhanced:
        i. The straight line estimation logic used to identify pass through volume can be refined, perhaps using optimization logic to identify the most likely routes a train would take
            1. This would allow the estimates to become more accurate and more granular
        ii. The risk scores assigned to commodities can be reviewed and refined
    b. The current risk score for commodity considers aggregate commodity volume; it's possible a more relevant metric would just consider the proportion of volume expected to consist of hazardous material
4. Identifying new data sources may encourage other ideas for refining existing or developing new features
5. Combined Risk Indices are currently assigned to track segments; this can be further refined
    a. Rail segments occasionally cross census tracts; there's likely an opportunity to separate the track segments for the purpose of more refined estimates
    b. Risk Indices can also be altered to extend outward from track segments (for example, a 1 mile radius using the East Palestine scenario as a reference)
6. Refine the weighting values used for the different components
7. The existing model can be expanded beyond Indiana to a national scope

    a. This should be plausible given that most data sources were provided by federal sources
8. The model generation pipeline can be cleaned and a combined ETL pipeline developed

Citations

AAR (Ed.). (2024, November 7). *Freight Rail State Data - AAR*. Association of American Railroads. https://www.aar.org/data-center/railroads-states/#:~:text=In%20a%20typical%20year%2C%20U.S.,nearly%20140%2C000%20miles%20of%20track.

Idem. (2023, February 24). *East Palestine, Ohio train derailment*. IDEM. https://www.in.gov/idem/featured-topics/east-palestine,-ohio-train-derailment/#:~:text=On%20Feb.,in%20manufacturing%20plastics%20and%20resins.

Schnoke, M., Lendel, I., Yochum, J., Driscoll, S., & Saneda, M. (2023, March). The Economic Consequences of the East Palestine Train ... https://engagedscholarship.csuohio.edu/cgi/viewcontent.cgi?article=2778&context=urban_facpub

U.S. Department of Health and Human Services. (n.d.). *East Palestine (Ohio) train derailment research response*. National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/research/programs/east_palestine