Optimising Rail Networks in California

Veronika Ulanova, Mark Pollock, Robert King, Hugo Hasted & Leon Wu

Word Count: 1580

January 17, 2019

Abstract

This report is aimed discuss, tackle and solve underdeveloped transport network problem in California, US. K-means clustering was used to find geographical areas of highly skilled workers with areas of high demand for employees. A minimum spanning tree algorithm and flow optimisation algorithm was used to find the ideal train route to connect workers to these jobs. The optimised map has been found to have a high resemblance to the high-speed rail project that is currently under construction.

1 Introduction

In California, cities with a population of over half a million are considered to be significant job-hubs, especially those near to San Francisco and Los Angeles. Research has shown that almost 14% of nation's venture capital investment is awarded to the Bay Area according to 2016 statistics [4]. Over the past 40 years the population of California has doubled, largely due to the well paid jobs in the technology sector encouraging people to move there. Unsurprisingly, this has driven the living costs up. The suburbs have become increasingly popular, with more people willing to commute further to work. Over the last 40 years California has not increased the capacity of it's transport systems at the same rate[4]. This is considered to be the primary reason for the slowdown in job growth in Silicon Valley. Despite that, there are still plenty of jobs with no easy means of getting there using public transport. This report aims to discuss this problem further as well as come up with a possible solution.

In this report, the 2010 US Census data [1], United States Cities Database [2] and Size of Business Data [3] was used to write an algorithm to optimise the potential national rail network. Using shapefiles also provided with the geographical census data were used to plot geographically accurate graphs, and provided the ability to draw county boundaries as well as state ones. The data set from the [?] was indexed by GEOID, with California split into 8000 regions called census tracts each containing a population of roughly 4000 people with a range from 1000 to 8000. The Census data provided over 70000 features such as employment status, education, household and commuting time. The goal was to optimise the rail network, this can be done in a myriad of different ways. The rail network could be optimised for connecting areas of high population density together; or connecting people with the longest commuting times or even connecting who earn the most, as there time is worth the most. The goal of building a rail network is to improve the community and create job growth, and it was felt that the best metric for this was connecting highly skilled and educated people, who currently are struggling to find work, to work. The features that were chosen to be analysed were therefore education attainment, employment status and jobs available.

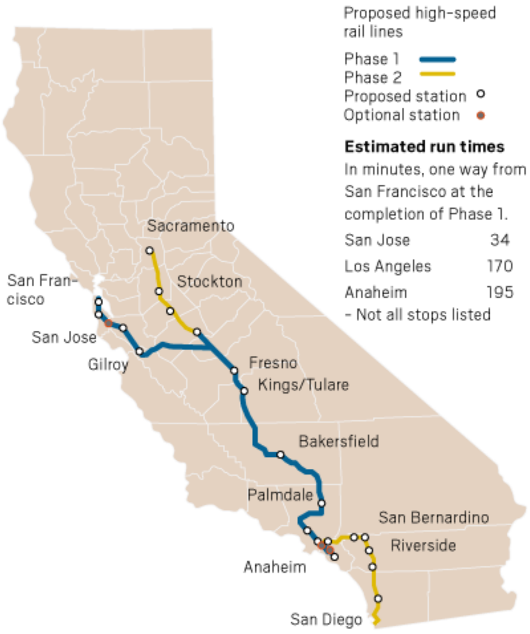

Figure 1: The map of the planned high speed California rail network

The performance of the optimised rail network map is judged by comparing it to the existing plans for the California High-Speed Rail project that is currently underway. The map of completed project shown on Figure 1 phase 1 is set to be finished in 2029, although there have been a number of mentioning of potential delays to this estimate due to unrealistic schedules [6] .

2 Method

2.1 Decision Tree

To gain data and understanding about the correlation between jobs and workers a decision tree model was trained of the US census data. The target feature was the 'Aggregate average income over the last 12 months' and two metafeatures were chosen: (1) the modes of transport and (2) duration of commute. The decision tree trained on 60% of the available data, retaining 20% for each of a validation and test dataset to minimise overfitting.

2.2 Cluster Plotting

In order to connect the wanted geographical areas discussed in Introduction map of California with a few of biggest cities by population were plotted. The clusters of highest number of: Figure 2 potential workers, and 3 jobs were formed and plotted as shown.

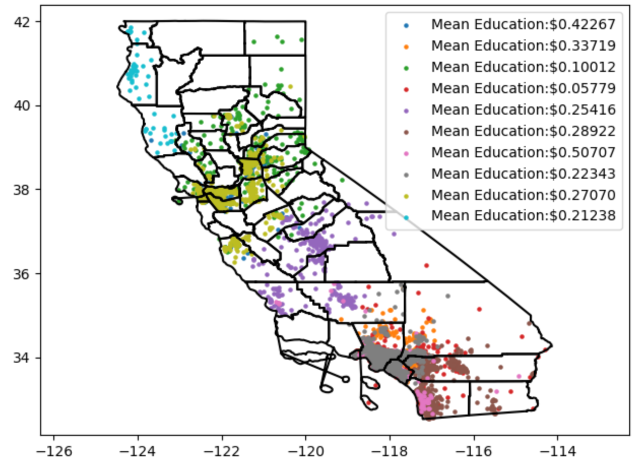

Figure 2: Shows 8 clusters corresponding to 8 areas of high numbers of educated potential employees.

Firstly, the two Census data sets had to be made consistent (see Introduction) to provide an estimate on the number of high skilled people per geographical area with the set of X, Y coordinates. This was acquired by matching the geographical identification code provided in each row of the US Census Geographical Data with corresponding number of educated people in that region from the US Census Survey [1]. That allowed for consistency between the map coordinates for every 4000 people region and the number of people with academic qualifications. The latter was given a weighting based on which qualification was acquired. If an individual just finished high school they were given a weighting of 0.5, which increased to 1 if they had completed a PhD. Moreover, the weighting was set to be sensitive to individuals that have dropped out of educational programs they enrolled on.Those were given less power in comparison to their counterparts that have finished the program. Figure 2 provides a visual map of clusters of high skilled workers in California.

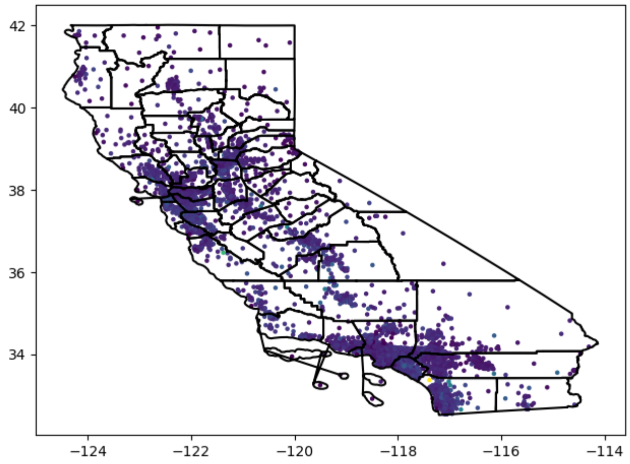

Figure 3: Shows a heatmap of the number of available jobs for each census tract. Warmer colors represent more jobs.

Similar was done to acquire the cluster map for the number of jobs available in the Bay Area as seen on Figure 3. For those the Size of Business Data was used, estimating the number of employee spaces. No weighting was performed on these data points as it was purely used to estimate the areas of employment for the nation rail network optimisation algorithm. The centers of each cluster were set to be nodes and used later as locations of the train stations. If clusters were within close proximity, 10km radius, they were removed iteratively.

[1] The advantage of clustering is that it massively reduces the number of nodes on the graph without losing precision in the data. The alternative approach is to make every data-point a node of the graph, but that would greatly increase the

2.3 Minimum Spanning Tree

The cluster map gives a measure of where regions of similar metrics are located. [1]. The centroid of each cluster position was identified, and this was added as a node to the graph. Using scipy the minimum spanning tree was found using the cartesian distance between nodes as the weighting of the graph. The minimum spanning tree minimises the total distance of the graph, which in the context minimises the total length of the railway network.

2.4 Flow Dynamics

In order to calculate the most in demand routes, the flow across each edge needs to be calculated. Each node was treated as a source or sink of employees (sources being residential districts, sinks being employers), with this the net flow along each edge could be calculated. An algorithm was built to do this which iterated through every node and balanced the flow in and out, using a method analogous to Kirchoff's Laws in electrical circuits. After acquiring our routes for different numbers of clusters, the flow dynamics are calculated within each system. This flow is calculated by considering the system as a whole, and computing the expected flow rate of workers towards jobs with respect to jobs/workers disparity and distance between stations.

3 Results & Discussion

The decision tree established a strong correlation between income and the chosen commuting features, with a accuracy, recall and precision of 0.82, 0.71, and 0.74 respectively. However, an issue with over-fitting needs to be resolved before this data can be used in other decisions and considerations.

Further optimasation could be performed. As seen in Figure 4, several solutions were produced, each using a different number of clusters.If number of clusters is increased, the total length and branching of the network increases as it connects more nodes. The optimal route based on the number of clusters could be established with an optimisation algorithm. A cost can be associated with the number of stations and the total length of the network. A gain can be associated with the number of people and number of jobs that have acquired access to the network, e.g. within 15km. The optimization could then be run to maximise the gain. In the algorithm a flow value was calculated for each edge. These edges could be used as a gain as part of an optimisation algorithm to remove links that do not provide more gain than their costs. Since the minimum spanning tree includes every node within the graph and does not consider their importance this could be a potential issue.

Figure 4 also shows the locations of the big cities in California, of population greater than 500 thousand. The locations of these large cities can be seen to lie on the proposed rail network. This validates that the network serves the areas of high populations and job opportunities. As the number of clusters are increased the rail network becomes longer and serves more rural communities. As the network becomes increasingly snake-like it becomes clear that the minimum spanning tree is not the optimum way to route between nodes.

Figure 1 shows California's proposed high speed rail network that is currently under construction as discussed in Introduction. A strong similarity can be seen between that route and our proposed route. Our proposed route shows more branching then California's planned route. This suggests that a longer rail network is more of an expense than our algorithm suggests, equally it could be said that the proposed rail network under-serves skilled, rural communities.

4 Conclusion

K-means clustering has proven to be an effective method to filter features using geographical location. The strong resemblance between the model and the planned infrastructure shows the success of the model. Comparison with the same algorithm implemented on population density would be valuable, to see if the more complicated optimisation for matching employees with employers had the desired effect.

References

[1] United States Census Bureau. *Geography. Maps and Data* Available from:`ftp://ftp2.census.gov/geo/tiger/` [Accessed: 18/11/2018]

[1] time taken to find the minimum spanning tree

[2] Simplemaps; Geographic Data Products. *United States Cities Database.* Available from: `https://simplemaps.com/data/us-cities` [Accessed: 18/11/2018]

[3] State of California Employment Development Department. State of Business Data - 2007 - present. *Number of Employees by Size Category - Classified by County (Table 3B).* Available from: `https://www.labormarketinfo.edd.ca.gov/LMID/Size_of_Business_Data.html` [Accessed: 17/11/2018]

[4] R. Mukherjee. *Silicon Valley Spotlight 2018* Available from: `http://blog.indeed.com/2018/01/18/silicon-valley-hiring-spotlight/` [Accessed:18/11/2018].

[5] K.Snibbe,Southern California News Group. The Mercury News. *The dream of high-speed rail in California is taking longer and costing more* Available from: `https://www.mercurynews.com/2017/03/14/the-dream-of-high-speed-rail-in-california-is-taking-longer-and-costing-more/` [Accessed: 18/11/2018]

[6] J. Daniels, CNBC: Politics. *California's $77 billion 'bullet train to nowhere' faces a murky future as political opposition ramps up* Available from: `https://www.cnbc.com/2018/03/12/californias-77-billion-high-speed-rail-project-is-in-trouble.html` [Accessed: 18/11/2018]

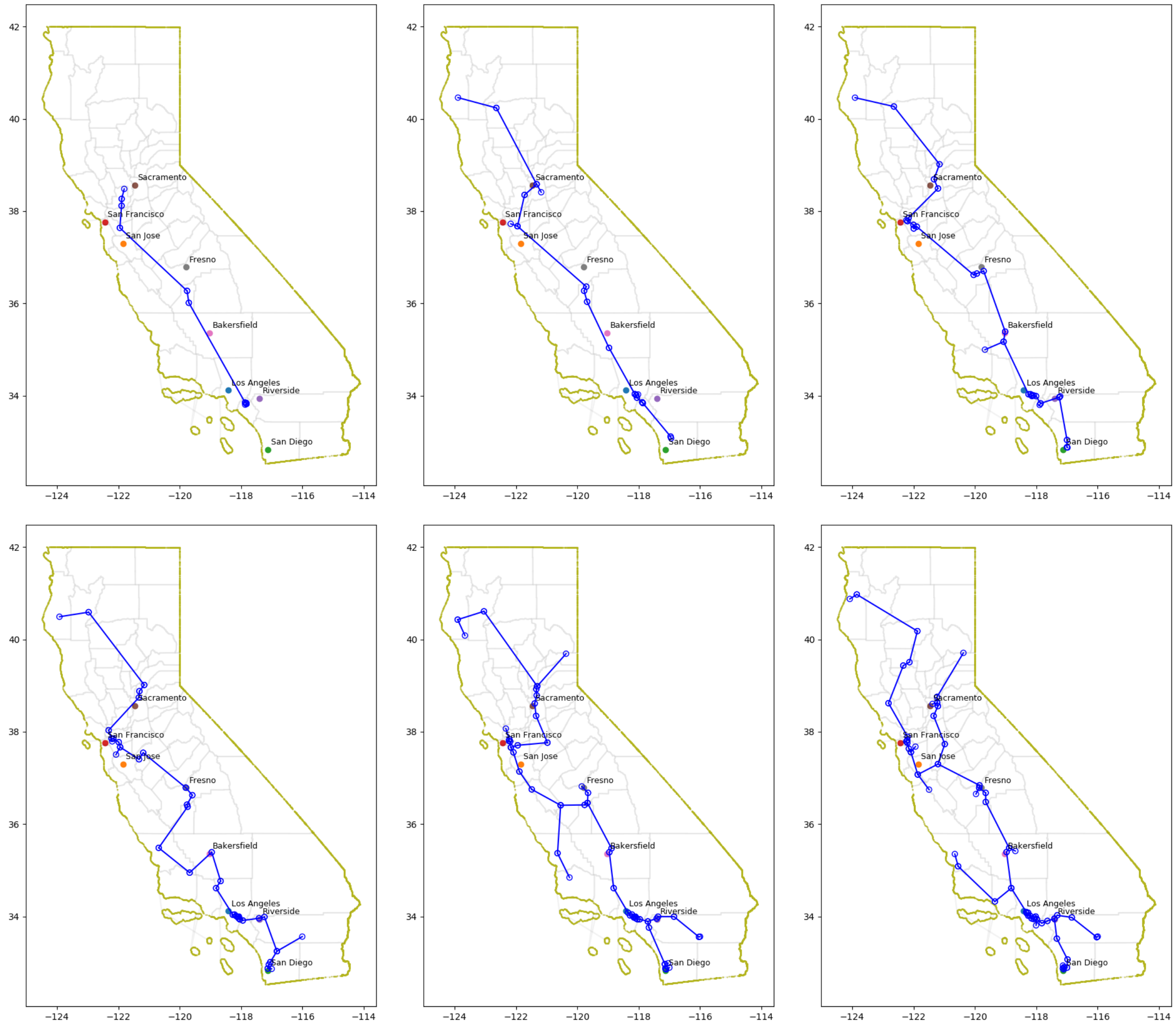

Figure 4: Figure shows 6 optimized rail network maps for various number of clusters used for node maps.