Mexico to U.S. Immigration

Brenner Swenson

Abstract— Utilising data from the Mexican Migration project and the American Community Survey, the project examines the development of Mexico to U.S. immigration over time using a visual analytics approach. The analysis focuses on the trajectories taken by illegal immigrants crossing to the U.S. and the effects of increased policies and border control on immigration volumes and coyote (smuggler) usage. Yielding that immigrants do not choose where to relocate solely for monetary reasons, this research solidifies the relationship between increasing border security and the rise in more complex and expensive border crossing operations.

1 Problem Statement

Mexican immigrants have long travelled to the United States to provide for their families in Mexico. It is commonplace for the head of household to immigrate to the U.S. to find seasonal work during the warmer months, where they then periodically remit money to Mexico until returning home in the Winter.

When Donald Trump became U.S. president in 2016, his administration introduced many policies directly targeted at reducing both legal and illegal immigration from Mexico to the United States. This research focuses on immigration trends, how they have changed over time, and the effects of Trump's presidency. Unfortunately, almost none of the data made available to the public can capture the effects of Trump's most recent 2018 policies, e.g. the Zero Tolerance Policy [1] or additional border security [2], as the most recent data was published prior to such policies.

Despite the lack of up-to-date data, there is still interest in analysing and quantifying the changes in immigration patterns over time. Using data from the Mexican Migration Project (MMP) and the American Community Survey Public Use Microdata Sample (PUMS), this research aims to answer the following questions:

1. Can visual analytics be used to gain an understanding of how and why immigration trajectories and methods have changed over time?

2. What factors can be observed in determining where an immigrant crosses the border, what jobs they are immigrating for, or where they immigrate to?

3. How have increased border security and the implementation of new federal policies affected immigration patterns and subsequently the composition of new immigrant inflows?

2 State of the Art

With the aim to obtain a better understanding of patterns in bird migration, Konzack et al. [3] tracked the data of 75 migrating gulls over a 3 year period using GPS data.

Their approach centred on the visual approach to their process, reiterating that ecological researchers are often slowed in their research by the technicalities of algorithms and coding. To mitigate these obstacles for ecologists and smooth the research process, a visual analytics tool was introduced that allows the user to view statistics of gull stopover sites, their overall trajectories, an interactive geographical map, a heatmap/calendar to filter the temporal component of the gulls' migration, and many more interactive features to pivot or filter the data on various properties. Konzack et al. specifically note the absence of trajectory aggregation in existing tools; citing that large volumes of trajectories without density aggregation impose visual clutter, thus hindering analysis. With regard to this analysis, the aggregation of individual trends for the visual analysis of Mexican immigrants is particularly applicable.

Konzack et al. ultimately presented their tool to an expert in the field of gull migration, providing the expert user with various analytical tasks such as "Find a stopover with a lot of gulls" (Stopovers are a break within a migratory trajectory), or "Which gulls are migrating from the breeding spot via North Spain to England?" [3] The user completed all tasks successfully, with the user confirming that their approach assists ecologists in exploring and visually identifying patterns prior to proceeding with more in-depth computational non-visual methods.

Using a geographical network approach, González Canché [4] analysed student migration patterns in the United States higher education network. Networks visualisations were overlaid on the U.S. map to display interstate relationships, quantifying the interdependence of supplier and receiver states in the context of non-resident student admissions to higher education institutions.

Also utilised was eigenvector centrality to quantify states' impact on both sending and receiving students to and from other states, a technique made famous by Google's PageRank Citation Index [5] to summarise every link on the internet into a single numeric value for prioritisation in Google searches.

The supplier eigenvalue centrality of any U.S. state grows relative to the state's output to universities that admit students from states that are highly active in the network. Conversely, receiver state eigenvalue centrality grows relative to the state's intake of students from states that contribute largely in the network and/or receive high volumes of students.

González Canché [4] found that resident students located in influential receiver states pay lower tuition in the public 2-year sector and instead pay lower tuition in the public 4-year sector, and that the amount of financial aid a high school student receives is inversely related to the likelihood they will leave their state of residence. These results encouraged states to increase admissions of resident applicants.

González Canché's findings are extremely relevant to the aims of this analysis; particularly relevant are the quantification of relationships between the origin and destination of a migration trajectory.

3 Properties of the Data

3.1 Mexican Migration Project Dataset (MMP)

The MMP was started in 1982 by the University of Guadalajara and Princeton University. The project focuses on collecting economic and social data on Mexico-to-U.S. legal and illegal migration.

Each year when migrants return to Mexico, random households are sampled in communities all over the country. Surveyors collect information on the economic status of the household, demographic information of its inhabitants, health data, and various social information. The heads of each household are queried on their most recent and first trips to the U.S., as well as a year-by-year account of the head of household's trips to and from the U.S.; these data include how they crossed the border, where they crossed, details on payment to so-called "coyotes" that assist in illegal border crossings, and can include particulars on up to 30 crossings per individual.

The MMP's data is partitioned across 7 core files, with each file containing information on different types of entities. The variables in each file are all numeric. However, the majority of numeric values in any given file are encodings of a categorical variable. For each file, these encodings are provided in a data codebook for the researcher to interpret. Each core file and their number of rows and columns can be found below.

File	Rows	Columns
PERS	176,701	132
MIG	8,823	612
MIGOTHER	939	244
CNMIG	65	145
HOUSE	28,331	548
LIFE	1,994,716	115
SPOUSE	814,885	85

Not all participants provide answers to each question. As such, each file contains null or unknown values which are encoded by the MMP as 8888 for null and 9999 for unknown. The amount of missing or unknown values for each core file can be seen in Figure 1. The high missing values in MIG are a result of the file containing space for up to 30 border crossings, where the median number of crossings per individual is 2. This causes the very high proportion of missing values.

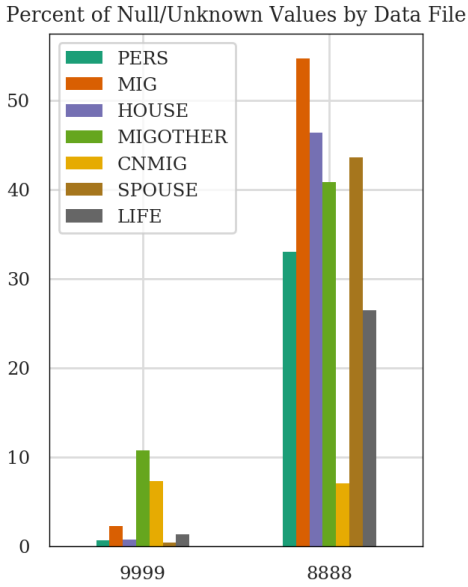

Figure 1. Missing and unknown values as a percent of all data points for each MMP file.

3.2 American Community Survey Public Use Microdata Sample (PUMS)

The ACS PUMS dataset is a sample of the American Community Survey intended for public use. The ACS is an ongoing survey targeted at even geographic coverage [6] where addresses country-wide have a 1-in-480 chance of selection for any given month, and the same address is not selected more than once every five years.

The files contain very similar demographic questions to the MMP data files, but instead are centred on the U.S. population.

Each annual file represents approximately 1% of the U.S. and thus allows end-users to estimate their findings to the general population.

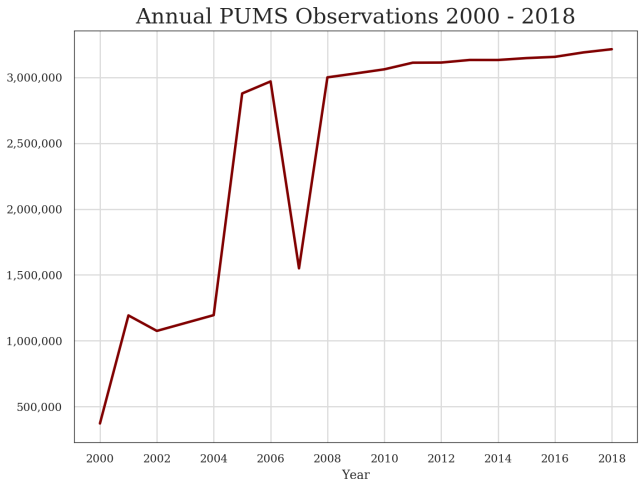

Figure 2. Sample size of PUMS data files over time. Large spike is present at 2007, most likely a consequence of the financial crisis.

The Census allows any user to download PUMS data since 2000. Each person-level file contains approximately 3 million records by 500 variables. After downloading and aggregating all 18 years of data, the dataset is 45.5 million rows. According to the World Bank [7], the U.S. population has grown steadily at a rate of approximately 1.5% from the year 2000 to 2018. Figure 2 displays growth in the PUMS sample sizes over time, displaying irregularities in 2007 when the U.S. population actually increased during that period. This is likely associated with the financial crisis and implies abnormalities in the PUMS data for that year.

4 Analysis

4.1 Approach

Beginning an analysis with Census-like demographic data is not an easy task as there are often many dimensions in the dataset. As seen in the table with the rows and columns of the MMP dataset, there are hundreds of columns to choose from, making it a challenge to form an initial starting point. Moreover, as research questions 1 and 2 are centred on geographical aspects of Mexican immigration, the geographic portion of the data transformation process will be the first step. The encoded geographic data in the MMP files will be decoded into the text representations of the location, but also encoded again with the latitude and longitude coordinates using a Python library.

Additionally, as the occupation codes in the data encoded with numeric descriptors, e.g. 410 for agricultural workers, the MMP appendix file containing the job titles for every immigrant mapped to their codes needs to be manually transcribed for decoding with Python. The MMP MIG file contains information on all border crossings of a migrant; the data is oriented as having a set of 6 columns for each crossing, contributing to the 612 columns observed above. Due to the orientation of the data, for time series analysis of border crossing data, the MIG file will subsequently need to be heavily transformed.

To begin answering questions 1 and 2, after basic descriptive statistics exploratory analysis, it will be most helpful to visualise the most common immigration paths for heads of household to understand where they are migrating most frequently regardless of time. After determining where migrants tend to travel to, it will be paramount to then investigate why they are going where they are going and attempting to do so with the variables present in the MMP file suite. After determining the possible reasons for why an immigrant might go to a certain state, a visualisation will help in determining the magnitude of the causal relationship between the destination of immigrants and certain variables.

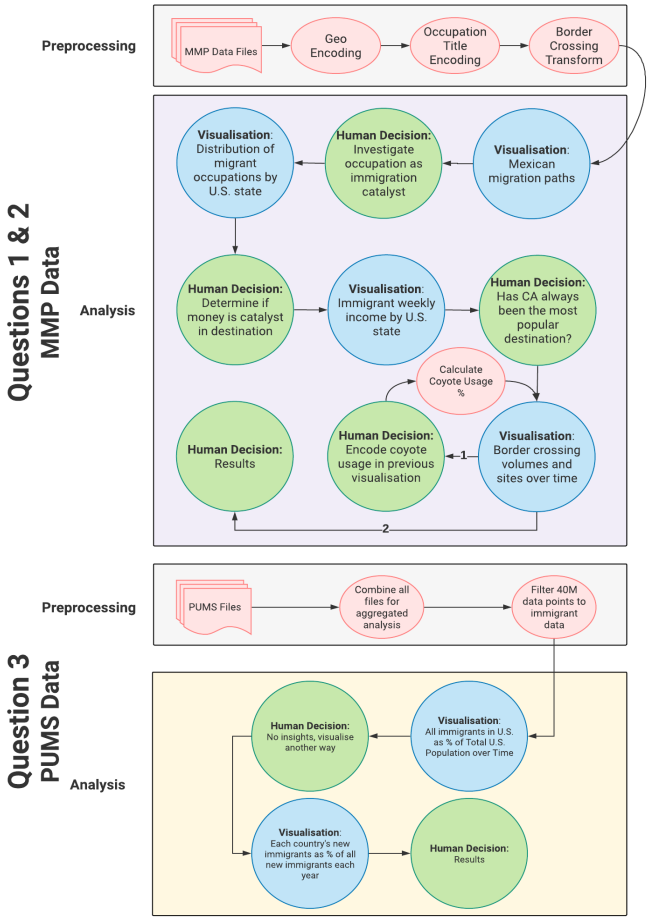

Figure 3. Resulting analysis workflow for Mexican immigration trends. Red indicates data processes and transformations, blue visualisations, and green human decisions.

For question 1 specifically, which is focused on development over time, the MIG file will be utilised as the border crossing data in said file is directly associated with a year, whereas in the PERS file only details on the first and last trip to the U.S. are present, which are not associated with a specific time period. After performing the data transformations previously mentioned, the border crossing sites will be geographically visualised with respect to time with additional variables encoded in the visualisation if possible, to help understand why the observed trends may be occurring.

To answer question 3, the PUMS dataset will be utilised as it contains information on the composition of the U.S. population. To analyse the dataset as a whole, all annual PUMS files will be downloaded and combined together to then be filtered to only the data relevant to immigrants. Subsequently, the volumes of inflows from various countries can be analysed over time using further visualisations.

4.2 Process

The MMP data required severe data cleaning and transformation prior to any analysis. For every file that contained location data, e.g. the state and municipality where a migrant was born or the state and municipality of where they crossed the border illegally, the latitude and longitude need to be encoded in order to visualise the geographic components of the data. Geopy [8] was used to encode the coordinates after converting the states and municipalities from their encoded 1990 FIPS codes [9].

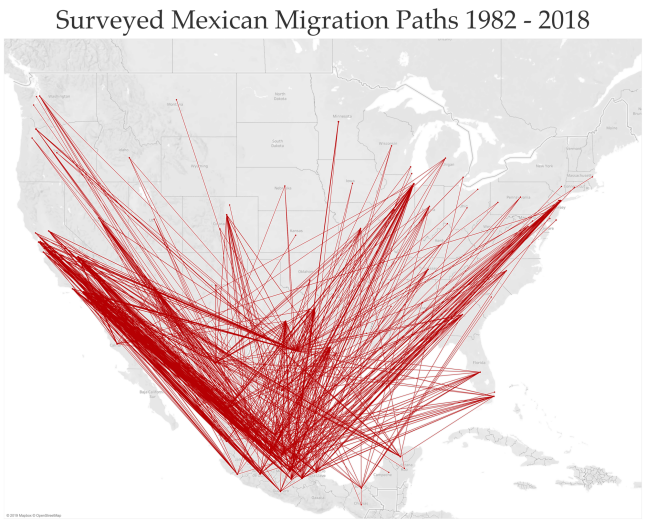

Figure 4. Mexican migrant trajectories with line thickness encoding number of trips.

The FIPS codes were very difficult to find online as the codes are constantly evolving as borders and jurisdictions change. The codes are hosted online by the Census but in a horrendous .txt file with inconsistent spacing; the codes could only be parsed using regular expressions, then merged back to the MMP data to successfully encode the names of the Mexican municipalities and states as they were in 1990 when the MMP used them for the first time.

The MMP LIFE file contains information on every year of a head of household's life since they were born. In addition to demographic information describing the individual, the file contains the state where they were born as well as the locations of the first and last U.S. location they migrated to for work. After applying the geo encoding process described above, all of the unique trajectories are displayed in Figure 4 with the thickness of the trajectory representing the number of times that specific path occurred in the dataset. Looking at Figure 4 one can observe a large volume of immigrants travelling to California; agricultural work comes to mind as the catalyst behind the trend. This trend warrants more analysis into what types of jobs are associated with each migration location. In order to contextualize the job titles, the occupation codes needed to be manually encoded from the MMP occupation codes appendices as mentioned in section 4.1.

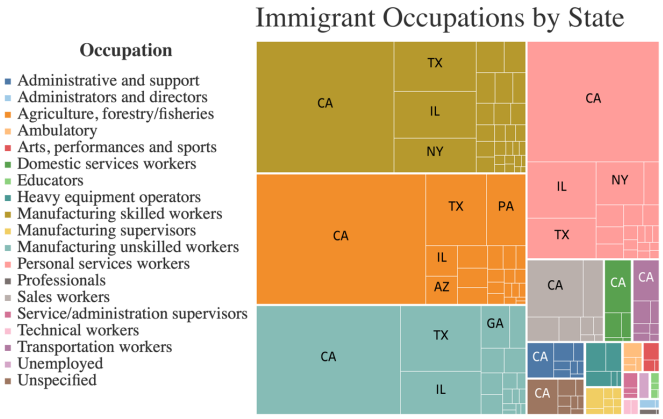

Figure 5. Immigrant occupations by state. A large majority of immigrants move to California for its agricultural and manufacturing jobs.

Furthermore, initial suppositions were correct. Figure 5 displays volume of immigrants per occupation title, with the state of the occupation broken out within the group. It is observed that the vast majority of migrants not only travel to California but are working in agricultural (orange) occupations. The more technical and academic jobs (e.g. technical workers or administrators) are much less prevalent as these roles most likely require extensive documentation and prerequisites for an applicant's consideration.

One would imagine that the majority of migrants would go to Texas as it is geographically closest and shares the largest border with Mexico, but that's clearly not the case. With such a high percentage of the surveyed migrants choosing California, what are the underlying reasons for doing so? The money? Ease of employment? Relatives that may already be living in the state? Higher chance of gaining legal residence?

This analysis can attempt to answer at least the monetary question by further transforming and modelling the MMP data as it contains data regarding the wages of an immigrant's job while they were living in the U.S. To provide an accurate representation of the majority of immigrants per state, the median wage by state can be calculated after removing egregious outliers.

The MMP file contains the hourly wage of each migrant, which on its own does not mean much as they could be paid a large amount hourly but only work one hour per week; consequently, the wage on its own could be misleading. Luckily the surveyors ask the respondent how many hours on average they would work per week and provided the number as an additional variable. This allowed for the derivation of the weekly pay for each migrant to allow for a uniform comparison by multiplying the two variables together.

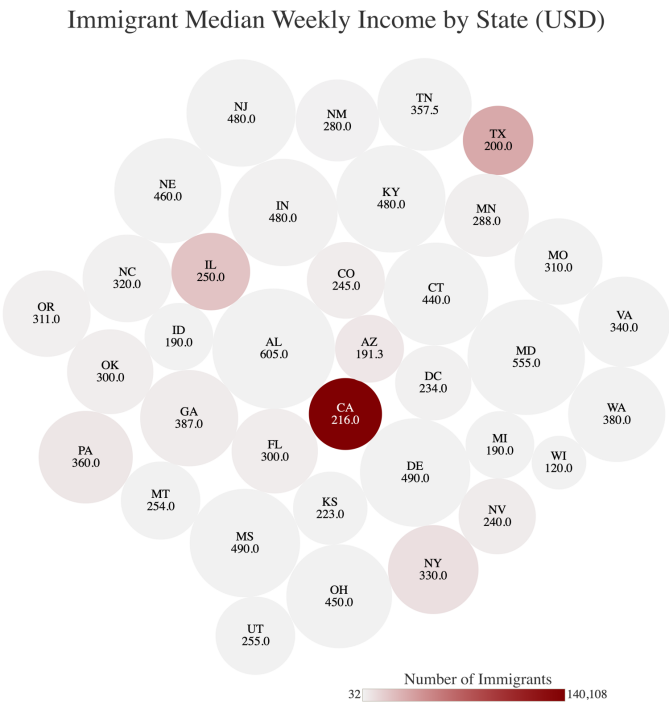

Figure 6. Median immigrant income encoded by bubble size with colour indicating the number of immigrants to each state.

With the radius of each circle encoding the median wage of each state and the shade of red encoding the volume of immigrants, Figure 6 illustrates that the wages of an immigrant's job are not a function of the number of individuals relocating to the job's state. In other words, immigrants are not relocating to California for higher-paying jobs relative to other U.S. states.

However, the immigrants that are traveling to states further away appear to be rewarded for their longer travels. This leads to an inference that it is much easier for an immigrant to find work in California than any other state. Alaska and Maryland have the highest median weekly salaries while also being two of the furthest states from Mexico. The prior analyses prove that California is by far the most favoured immigration destination among Mexicans, but has that always been the case? Figure 4 displays the trajectories taken from the MMP LIFE file; these paths are a migrant's last trip to the U.S. as of the survey date. Consequently, this means that the observed paths are not associated to any specific year and can subsequently not be used as a dimension for further analysis.

The MIG file, however, contains detailed information on individual border crossings including but not limited to: where they crossed, how they crossed, health data, finances, and most importantly when they crossed. Each row in the MIG file represents a person and contains information up to 30 border crossings per individual with metadata for each crossing. To prepare the border crossing data for visualisation, each crossing needed to be parsed from each row while still maintaining the other metadata. This complicated operation transformed the data from 8,823 observations by 612 variables to 16,943 by 16 after removing all null border crossings.

When choosing to cross the U.S. border, an immigrant is placing themselves at very high risk for apprehension by the U.S. Border Patrol. Since the early 2000s nearly 700 miles of physical barriers have been established across the Southwest U.S. Mexico border [10] in addition to the implementation of security technologies such as thermal imaging, radiation portal monitors, ground sensors, aerial drones, and license plate readers. [11] Currently, as an outgrowth of current policies, children crossing the border with their parents are being detained in inhumane conditions for months on end while their parents are jailed and prosecuted. [12]

As a result of such measures, it has become harder and harder for immigrants to successfully cross into the U.S. without proper documentation and thus deterring immigrants from attempting a trip, or forcing those who want to enter the country to do so with assistance via costly and increasingly complex smuggling operations.

Those who provide border smuggling services are colloquially known as ‘coyotes’. Individuals or full families can pay up to $15,000 for their assistance in crossing the border; this is a steep increase from where it used to be in the early 2000s (approximately $200), but due to increasing security, the prices have skyrocketed in recent years [13].

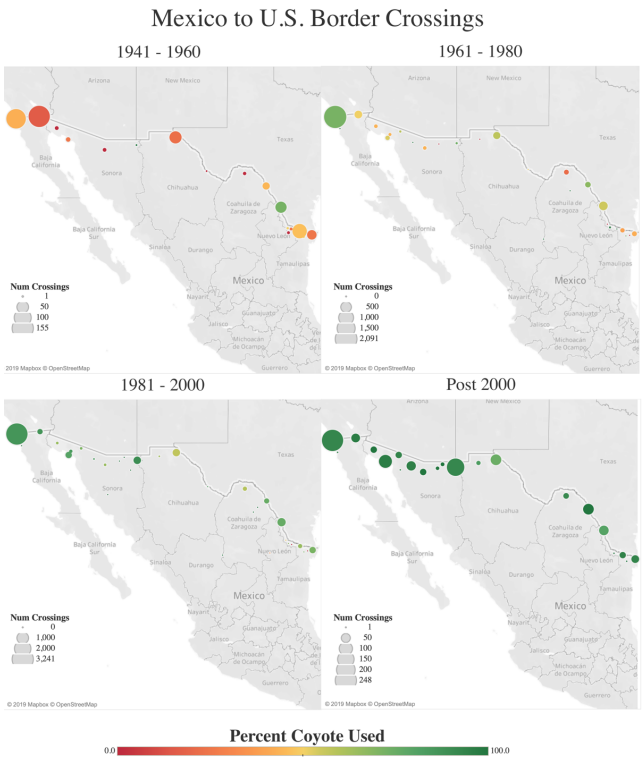

Figure 7. Mexico to U.S. illegal border crossing locations over time with percentage of coyote usage encoded as the point's color and the number of crossings as the point's radius.

The head of household surveyed in the MMP MIG file is asked whether or not they used a coyote in crossing the border, allowing insights to be extracted regarding the development of coyote use over time. Figure 7 (above) illustrates the number of surveyed crossings from 1941 to 2017 with the data partitioned in to 20-year groups. The aggregations of the crossing locations, that were geocoded, are displayed as a circle whose radius indicates the number of crossings at each location. Furthermore, the colour of each circle encodes the percentage of border crossings that used a coyote at each location.

It is observed that coyote use has increased drastically over the past 50 years. It is possible to infer that Tijuana, the largest circle in each sub plot, remains to be the most popular place to cross the border. However, the number of crossings appears to decrease dramatically post 2000, which may be a result of the survey methodology, but also most likely holds statistical relevance as the number of apprehensions by the U.S. Border Patrol has decreased by nearly 82% since FY 2000. [14] This observation leads to a further question: how has the Mexican immigrant representation within the U.S. population changed with time?

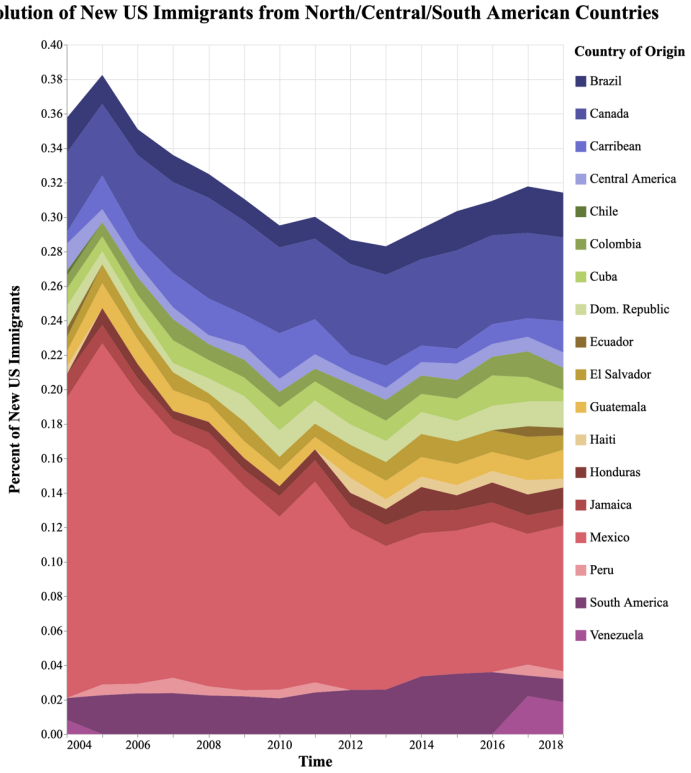

Figure 8. A stream chart exhibiting the composition of the new U.S. immigrant population over time. Observed is a large decrease in the % contributed by Mexico.

The PUMS dataset, as mentioned above, contains information on the composition of the American population, and is published annually. Due to the volatility in the sample sizes as evidenced in Figure 2, it would be naïve to analyse the discrete counts of surveyed U.S. residents. Instead, the representations year on year as a percentage of some denominator is preferable as it is not subject to sample size.

Initially the number of immigrants per country were compared to the total number of self-identified immigrants year on year, but due to the sheer volume of immigrants in the U.S. the results were not informative. Instead, as seen in Figure 7, it is more enlightening to visualise only the survey respondents who identified as living in another country a year prior to when they were surveyed. Thereafter comparing their country of origin to the rest of the respondents who identified as living outside the U.S. one year ago.

4.3 Results

To answer questions 1 & 3, using an iterative visual analytics approach, the border crossing data was explored using variable permutation to determine that coyote usage has increased significantly with time. Additionally, coyote usage increased as the amount of security on the border increased as seen in Figure 9. The number of apprehensions [15] on the South West border are visualised against the number of Border Patrol employees [16] as well as coyote usage percentage over time.

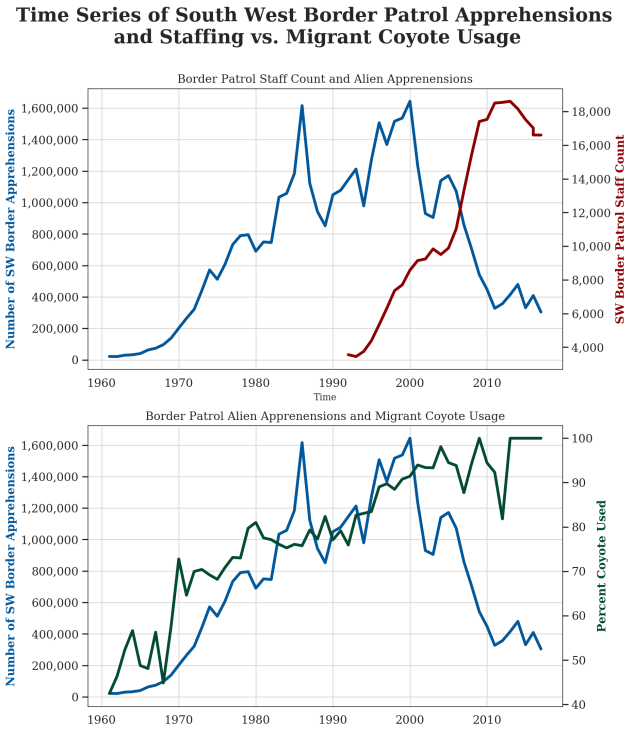

Figure 9. South Western Border Patrol apprehensions and staffing over time. Bottom plot displays percentage of coyote usage.

There is a clear relationship between the increase in Border Patrol and the increase in coyote usage; there is also an inverse relationship with the number of Border Patrol staff and apprehensions; this implies the increased security serves as a deterrent to would-be migrants. The PUMS data in Figure 8 shows a large decrease in % of new Mexican immigrants at the same time as Border Patrol staff count increases drastically.

For question 2, immigrants choose to travel to California far more than any other state. They choose primarily agricultural and maintenance occupations. Additionally, they do not travel to CA for the money, as the median weekly income is fairly low when compared to other states. Immigrants that travel to states further from Mexico tend to get paid more than their CA counterparts.

5 Critical Reflection

Analyzing very macro-level data involving populations and survey data is a difficult task as often the variable relationships are obfuscated due to surveying methodologies. The approach used in this analysis was entirely dependent on the assumption that the underlying data was clean. To that end, for future researchers using the MMP dataset, it would be beneficial to perhaps communicate with MMP to further understand the inconsistencies and peculiarities in the data.

Unfortunately, due to the nature of the dataset and the domain, only the municipalities of the geographical data were available. This then led to all data points associated with one municipality to be geocoded to one specific geographical coordinate. If the exact latitude and longitude of crossings, job locations, etc. were available in the dataset, then further geo-spatial clustering could have been performed to further understand the people that migrate to certain places.

Moreover, the visualisation of immigration trajectory density found in Figure 4 proved to be extremely fruitful in determining which states to further analyse, e.g. California. Another limitation of the dataset was the absence of any intermediary point in migration, i.e. a stopover location where an immigrant might stay the night prior to crossing a border. Such information could provide incredible insight for further trajectory analysis, especially combined with coyote data as the researcher could potentially understand how smugglers operate vs those who attempt to cross the border on their own.

Human intuition played a large part in the interpretation of results; the choice to investigate occupations and wages as catalysts for immigration is a result of the researcher having lived in an area of the U.S. with large numbers of Mexican immigrants. Subsequently the researcher is very familiar with the motivations of many immigrants, and the drive to earn money to remit back to Mexico to support their families. The analysis did not factor in any remittance variables in the dataset, even though they were present. Future researchers can investigate the relationships between geographical locations in the U.S. and who is sending the most money back to Mexico.

The analysis process was based almost entirely on the outcomes of various graphs, which was quite suitable for this domain and set of problem statements. Being able to visualise the data geographically, internalise the results, then pivot the visualisations to investigate another variable produced insights that otherwise would not have materialized if the analysis was performed only computationally.

The use of visualisation software such as Tableau, as opposed to Python, proved to be very beneficial and efficient. For exploratory analysis using visual methodologies, the ability to replace variables and pivot charts within seconds subtracted what would have been hours from the analytical process. For future researchers utilizing Census-like data, it is recommended to use visualisation software to avoid spending copious amounts of time coding complicated plots to only then re-iterate and attempt another plot upon finding little or no results.

6 Table of word counts

Section	Word Count	Limit
Problem statement	258	250
State of the art	499	500
Properties of the data	493	500
Analysis: Approach	487	500
Analysis: Process	1478	1500
Analysis: Results	210	200
Critical reflection	482	500

References

[1] 'Attorney General Sessions Delivers Remarks Discussing the Immigration Enforcement Actions of the Trump Administration', 07-May-2018. [Online]. Available: https://www.justice.gov/opa/speech/attorney-general-sessions-delivers-remarks-discussing-immigration-enforcement-actions. [Accessed: 30-Nov-2019].

[2] 'More Troops Deploy to Support DHS/CBP Southwest Border Mission > U.S. DEPARTMENT OF DEFENSE > Story'. [Online]. Available: https://www.defense.gov/explore/story/Article/1675862/more-troops-deploy-to-support-dhscbp-southwest-border-mission/. [Accessed: 30-Nov-2019].

[3] 'Visual exploration of migration patterns in gull data - Maximilian Konzack, Pieter Gijsbers, Ferry Timmers, Emiel van Loon, Michel A Westenberg, Kevin Buchin, 2019'. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/1473871617751245. [Accessed: 30-Nov-2019].

[4] M. S. G. Canché, 'Geographical Network Analysis and Spatial Econometrics as Tools to Enhance Our Understanding of Student Migration Patterns and Benefits in the U.S. Higher Education Network', *Rev. High. Educ.*, vol. 41, no. 2, pp. 169–216, Dec. 2017.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. 1998.

[6] U. C. Bureau, 'American Community Survey Information Guide', *The United States Census Bureau*. [Online]. Available: https://www.Census.gov/programs-surveys/acs/about/information-guide.html. [Accessed: 15-Dec-2019].

[7] 'Population, total - United States | Data'. [Online]. Available: https://data.worldbank.org/indicator/SP.POP.TOTL?end=2018&locations=US&start=2000&view=chart. [Accessed: 15-Dec-2019].

[8] 'GeoPy 1.20.0 documentation'. [Online]. Available: https://geopy.readthedocs.io/en/stable/. [Accessed: 17-Dec-2019].

[9] Census Bureau, 'METROPOLITAN AREAS AND COMPONENTS, 1999, WITH FIPS CODES', *Census.gov*. [Online]. Available: https://www.Census.gov/population/estimates/metro-city/99mfips.txt. [Accessed: 17-Dec-2019].

[10] J. G. J. Gelatt, 'President Signs DHS Appropriations and Secure Fence Act, New Detainee Bill Has Repercussions for Noncitizens', *migrationpolicy.org*, 01-Nov-2006. [Online]. Available: https://www.migrationpolicy.org/article/president-signs-dhs-appropriations-and-secure-fence-act-new-detainee-bill-has-repercussions. [Accessed: 20-Dec-2019].

[11] U. S. G. A. Office, 'Southwest Border Security: Border Patrol Is Deploying Surveillance Technologies but Needs to Improve Data Quality and Assess Effectiveness', no. GAO-18-119, Nov. 2017.

[12] 'Trump administration detains nearly 70,000 migrant children in record high', *The Independent*, 12-Nov-2019. [Online]. Available: https://www.independent.co.uk/news/world/americas/us-politics/trump-us-mexico-border-immigration-detention-children-family-separations-a9199686.html. [Accessed: 20-Dec-2019].

[13] J. Rohrlich, 'This is how much it costs to be smuggled over the US border', *Quartz*. [Online]. Available: https://qz.com/1632508/this-is-how-much-it-costs-to-cross-the-us-mexico-border-illegally/. [Accessed: 20-Dec-2019].

[14] 'Border Security Along the Southwest Border: Fact Sheet', *National Immigration Forum*. [Online]. Available: https://immigrationforum.org/article/border-security-along-the-southwest-border-fact-sheet/. [Accessed: 20-Dec-2019].

[15] 'Stats and Summaries | U.S. Customs and Border Protection'. [Online]. Available: https://www.cbp.gov/newsroom/media-resources/stats. [Accessed: 21-Dec-2019].

[16] 'U.S. Border Patrol Fiscal Year Staffing Statistics (FY 1992 - FY 2018) | U.S. Customs and Border Protection'. [Online]. Available: https://www.cbp.gov/document/stats/us-border-patrol-fiscal-year-staffing-statistics-fy-1992-fy-2018. [Accessed: 21-Dec-2019].