Grace Gao

Bass Connections

November 20th, 2020

The Art of War and Migration: Semester Report

**Introduction**

When we think of migration, we think of patterns. The patterns of movement from one place to another. But what exactly causes migration? Often, the reason for migration is subtle and difficult to pinpoint an exact cause. Other times, the cause is blaring obvious. However, as a researcher, it’s my job to use data to analyze and predict what these possible reasons could be. For this Bass Connections project, our topic is the Migration, Urban Cultures and the Arts. In particular, the goal is to develop a data-driven study of migration and its relationship to arts and culture. As an enthusiast for European art, I really wanted my project to remain euro-centric while delving deep into the past. Thus, for my research project, I am analyzing the effects of the Eighty Year’s War on the migration pattern of artists to and away from Antwerp and Amsterdam, two large cultural centers of Netherlands at that time.

So, what exactly was the Eighty Year’s War? The Eighty Year’s War (also known as the Dutch War of Independence 1568-1648) was a revolt of the Seventeen Provinces of the Netherlands again the Spanish Crown. With a decent-sized population of Protestants, mainly the branch of Calvinism, in Netherlands and the Iconoclasm, King Phillip II’s sent troops to violently suppress Protestants. In addition to centralized rule, different cultures, taxes, and competing trades, the Dutch began to revolt, which ultimately led to the span of the Eighty Year’s War and the eventual recognition of the Dutch Independence. During this war, a few notable events would be the Fall of Antwerp (August 17th, 1585), The Twelve Year’s Truce (1609), and Peace of Munster (January 30th, 1648). The Fall of Antwerp was when Antwerp fell to the Spanish army after a siege that last over a year. At that time, Antwerp was the largest city and the cultural economic, financial center of the seventeen provinces and in November of 1576, the Spanish soldiers revolted and burned the city during the Spanish Fury. This led Antwerp to become even more engaged in the revolt, becoming the capital of the Dutch Revolt. After the fall, the Protestants were given four years to settle their affairs before leaving the city. The Twelve Year’s Truce stopped hostilities between the Hapsburg rulers of Spain, southern Netherlands and the Dutch Republic. Finally, the Peace of Munster formally recognized the Dutch Republic.

I chose this topic mainly for two reasons. First, the long-time span of the war allows for a larger dataset for analysis and the concept of war is very life altering to anyone and everyone living in the affected areas. Specifically, with the multiple triggers that started the war (religious persecution, economic hardships, etc) and the fear, violence and bloodshed that followed, it provides numerous incentives and variables to analyze as the cause of migration. However, my analysis isn’t just restricted within the eighty years of the war. My dataset will also include datapoints that encompasses the pre-war period and the post war-period to fully account for any possibilities or subtleties in regard to the migration trend. But the real analysis, comes from implementing machine learning models onto the dataset to statistically predict these patterns of migration.

Unfortunately, the Random Forest Regression Model, which I hope to utilize, is not very common among historical data. For my research, I consulted two studies that implemented this specific model. One was a paper “One Millisecond Face Alignment with an Ensemble of Regression Trees” written by Vahid Kazemi and Josephine Sullivan, which discussed the implementation of regression trees to estimate the face’s landmark positions based on an image. In particular, they presented the algorithm’s ability to perform face alignment in milliseconds. With the cascading regression functions, they were able to efficiently estimate the shape of the face from an initial estimate and indexing of the pixel intensities from the initial estimate. For the second study, it was conducted by Gunjan Mishra, Deepak Sehgal and Jayaraman K Valadi, who utilized the Random Forest and Extra-trees regressors to predict the anti-hepatitis activity of antimicrobial peptides, which are host defense peptides that could potentially be used to replace broad-spectrum antibiotics. For this study, they collected data through publically available resources and actually collected 8809 descriptors (identifying features) but cut it to the 157 descriptors that were the most applicable. In the face alignment study, the data used was essentially all numerical and missing data was not an issue. With numerical data, it’s a lot easier for models to analyze and create an accurate prediction. Similarly, with the medical study, a large portion of the dataset was numeric, and it contained 157 predictors allowing for a very comprehensive analysis. In addition to the enormous size of the dataset, it didn’t really encounter missing data. However, working with historical data, missing data is extremely common since not every detail can be accounted for from hundreds of years ago. Furthermore, the accuracy of the historical data could also be questionable. Fortunately, the database I’m relying on seems very comprehensive. Finally, a lot of the data is categorical rather than numerical. With categorical data, it becomes a lot more difficult for the model to accurately predict a trend.

Although for my project I will using the same machine learning model—Random Tree Regression—as the studies that I consulted, the analysis of the historical art data really differentiates itself in this machine learning field. In addition, the mixture of categorical and numerical data makes my project more complicated as opposed to the datasets for these studies. Regardless of the differences, I’m still able to draw inspiration from these studies and alter their approaches to best fit my project.

**Research Question**

Knowing the background and a brief overview of my approach, this study hopes to answer the following research questions through analyzing the historical art data scraped from the ECARTICO database:

1. How does the background of an artist—nationality, religion, experience, age, gender, occupation and family size—influence their decision of migration?
2. How does the artist migration model differ between Antwerp and Amsterdam? Specifically, 1556-1568 (When Spain conquered Netherlands, but before the Eight Year’s War, 1568-1648 (Eighty Year’s War), and 1648-1675 (Golden Age/Post-War).

With these research questions, I have my own hypothesis based from the current research that I have done.

1. I believe the background of an artist will largely influence their migration decision. However, the only identifying feature I’m confident in making a prediction about is religion. In particular, those who are Protestant will migrate outside the war’s boundaries to escape the religious persecution. For the other variables, I think the prediction generated from the model will really surprise me. Often, when working with data, the statistical analysis uncovers trends from seemingly unrelated data or disproving popular perceptions.
2. For this portion, I believe prior to the Fall of Antwerp, we will see a larger migration to Antwerp rather than to Amsterdam. However, there will still be a constant stream of migration to Amsterdam. However, after the Fall of Antwerp, we will see a huge shift of the migration numbers to Amsterdam and an exodus away from Antwerp. Overall, I don’t think there will be a notable migration away from Amsterdam throughout the time periods analyzed. Eventually, I hope to extend my analysis beyond just these two cities. However, I hope to do this once I’ve finished the analysis for these two cities.

**Methodology**

My project is completely data centric, which means I have to have access to an analyzable dataset. However, this isn’t as simple as it may seem. In order to have this dataset, I must web scrape from the ECARTICO Database. Originally, I had hoped through extracting the API, I would be able to access all the data that I needed. Unfortunately, the API was not very comprehensive, which meant I had to scrape directly from the database. In order to this, I will be using the language R and the rvest package. First, I will scrape all the names and specific URL links to each individual’s page and formulate this information into a dataset. This means my dataset will contain all the datapoints that ECARTICO has. Next, I will utilize each individual URL to access each artists’ specific page. From this page, I will scrape the necessary variables and combine all the information into one dataset. In this dataset, each artist will have his/her own row with the variables as the columns. From this dataset, I will filter by the time periods and locations to gather all the datapoints for my specific needs. Throughout this process, I will have to clean and format the data so it will the most accessible when implement the machine learning model.

To truly understand the implementation of this machine learning model—Random Tree Regression--and the application of it to my project, I need to dive into the methodology of the algorithm. The general concept of random forests classification was proposed by Leo Breiman with the specific instance of the RF-CART. Although I won’t dive specifically into the history of the algorithm, I will be discussing the actual logic behind it. As is its name suggests, a regression tree is built by recursively splitting the data into more and more homogenous “nodes” until it reaches the “terminal node” (Grömping 4). Each split is based on the division of a variable, which is choose by a “splitting criterion”. To conduct the predictions, you follow along a branch of the tree until the specific terminal node. With this specific path, you can find the predicted response value, which is just the specifical terminal node’s average response. For the random aspect of the random forests, each tree is built from a random subset of the data, and each split is based randomly on a subset of the different potential variables. Although the randomness causes varying predications in each tree, the averaging of these predictions’ accounts for the variance.

Overall, this algorithm is ideal as it’s able to be implemented in a large set of data (like the ones I will use) and obtain non-linear relationships among the variables. Furthermore, it’s more favorable than statistical models like the linear regression model or forwards/backwards elimination as it’s able to account for categorical variables more accurately than the linear regression model could. Particularly for my data, the only numerical variables are age, the family member count, and the different years. So, it’s crucial I have a good algorithm that can account for the numerous categorical variables and still make an accurate prediction. Finally, the randomness and the averaging behind the algorithm enables a more accurate prediction than other models, especially since many of my variables could be more correlation rather than causation.

**Data Description**

To achieve my dataset, I will be web scraping from the ECARTICO database, which is a comprehensive collection of structured biographical data of artists. In the database itself, it has biographical data on 53,248 people. During my web scraping process, I will actually be collecting all the data points and then filtering it to fit for my time period and location. However, currently I do not have an exact number of entries for my final dataset because of the complicated nature of web scrapping and cleaning the data.

But, in my dataset, I will have 21 variables: six dummy, seven text and eight numerical. The dummy variables are gender, pre-war period, war period, post war period, 50 years prior to the pre-war period and 50 years after the post war period. For the gender variable, it will be one if the artist is female and zero if the artist is male. For the time periods, it will be one if they are alive during the period, which means the artists could be alive throughout multiple time periods. The purpose of these time periods is to determine exactly when the migration occurred relative to the actual span of the war. It also accounts for the life span of the artists. Finally, I want to see how gender plays a part into migration. Because of certain societal norms and stereotypes, I would think that there would be a small proportion of female artists. In addition, from this small proportion, I think that an even smaller proportion would actually migrate. However, this is only a perception that could easily be proven wrong from the actual data analysis.

The text variables are birthplace, death place, father name, mother name, occupation type, location and religion. Although the father and mother name may not be the most applicable, but it still accounts for the artists who have a recorded parent’s names. However, for the occupation type and location, they will be a vector of text variables since many artists have several different occupations and have resided in multiple locations. Since these variables are in a vector form, it’s important to find ways to split these vectors to analyze individually each value. The purpose of this location variable is to really understand the pattern of migration. For some artists, they could simply be residing in an area for one to two years as opposed to 10-15 years or the rest of their lives. In addition, it will be interesting to see how these artists actually move. Do they move to a certain area during the war and move back to their birthplace after it? Do these artists travel great lengths to escape the war’s boundary to a foreign place with a different culture and language? These are all the questions I hope to answer with this variable and the corresponding dates variables in the numeric section. But, with any historical data comes the missing data. Unfortunately, from a preliminary glance, it seems that the religion variable will contain a lot of missing information. This is really tragic since a main cause of the war was religious persecution. So, it’s important to consider how those who identified with a certain faith migrated during this whole time period. However, hopefully there will still be a decent size of entries that contain this religion variable. Finally, for the birthplace, it’ll allow me to determine the nationality of the artists while analyzing where exactly most artists originated from prior to any migration. In addition, the death place will allow me to analyze the end point of migration and discern if there was a pattern in where artists lived at the end of their time.

Finally, the numerical variables are birth and death dates, spouses and children count, location start and end years, column number and artist name. Specifically, for the artist name, I will most likely end up assigning a numerical barcode to the artist rather than maintaining their name throughout the analysis part. For the birth and death dates, it is to determine which time periods their life spans fall under. However, I may choose to add an additional column of just the years rather than the specific dates because of missing data. For the location start and end years, these variables will be numeric vectors. Each pair of years will then correspond to a location in the location text vector. For the spouses and children count, I will be using these variables to analyze if family size factors into an artist’s motivation to migrate.

For my selection, the data will range from 1556-1725. So, this time period serves as the first filter for my dataset. Next, anyone born, died, or lived in either Amsterdam or Antwerp during this time period whose life span falls within this time period will be included in my dataset. However, I do believe I will eventually expand this location criteria beyond just these two cities to the most prominent cities of Netherlands at that time.

**Preliminary Analysis and Visualizations**

Unfortunately, I don’t have any visualizations just yet. The process of web scraping is very tedious and tidying the dataset is also very time consuming, which means I can’t create any visualizations until I’ve extracted, cleaned and tidied the dataset. However, I hope to eventually use Tableau to create very appealing and pleasing visualizations of the migration pattern. Particularly, I want to actually display this pattern onto an actual map that will be interactive. Finally, I want to try and develop a website using HTML/CSS and JavaScript that will display all my visualizations, my dataset and my statistical analysis.

**Next Steps**

As the semester ends, it’s important to lay out the next steps for this project. During this semester, I was able to solidify my research project topic, consult different studies, and begin the web scraping process. However, I wasn’t able to compile my dataset this semester. So, first, I want to finish web scraping, cleaning and tidying the data by the end of winter break. Starting next semester, I hope to have a completed and accessible dataset to start implementing the regression model. In addition, I want to find more studies that have used the Random Forrest Regression and learn more about their approaches and all the nuances that were applied to fit their certain topic. Although currently I am set on using the Random Forrest Regression, I actually want to expand beyond just this model. So, I hope to do more research next semester in finding additional statistical models to use. With more models, I can compare the different predictions and find the most accurate analysis. So, on the model side, I want to research more studies that have implemented machine learning models to apply to my own.

Aside from completing the dataset and conducting the statistical analysis, I want to really work on the visualizations. This means I want to focus more on using Tableau or other visualization tools to really map out all this data. As mentioned before, I think it’d be very interesting to actually display the patterns of migration on an actual map of that time to visualize the pattern. In addition, I want to try and make it interactive, so users can see the change over time of the migration pattern throughout this time period.

Finally, I want to add all this analysis, visualizations, dataset and conclusion onto a website that I will create using HTML/CSS and JavaScript. With this website, it will be compiling all this information into an accessible format that anyone can navigate. Although I have a basic knowledge on building this website, I think I would need more help in learning bootstrap, which I’ve heard is very useful in creating a visually pleasing website. In addition, there are a lot of display formats I would like to implement with this website but may need further guidance in going about doing this.

Overall, I am very excited in the next steps of this project. I believe the completely data approach and the statistically modeling of this data is very unique within this topic. This project has allowed me to combine my passion of art and data and I am ecstatic to see the final findings and the end product.

**Works Cited:**

*ECARTICO: Linking Cultural Industries in the Early Modern Low Countries, Ca. 1475 - Ca. 1725*, www.vondel.humanities.uva.nl/ecartico/persons/.

Grömping, Ulrike. “Variable Importance Assessment in Regression: Linear Regression versus Random Forest.” *The American Statistician*, vol. 63, no. 4, 2009, pp. 308–319., doi:10.1198/tast.2009.08199.

Kazemi, Vahid, and Josephine Sullivan. “One Millisecond Face Alignment with an Ensemble of Regression Trees.” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, doi:10.1109/cvpr.2014.241.

Mishra, Gunjan, et al. “Quantitative Structure Activity Relationship Study of the Anti-Hepatitis Peptides Employing Random Forest and Extra Tree Regressors.” *Bioinformation*, vol. 13, no. 03, 2017, pp. 60–62., doi:10.6026/97320630013060.

Nijboer, Harm, et al. “The Painting Industries of Antwerp and Amsterdam, 1500−1700: A Data Perspective.” *Arts*, vol. 8, no. 3, 2019, p. 77., doi:10.3390/arts8030077.

Riddell, Jennifer. *Painting in the Dutch Golden Age: Classroom Guide*. National Gallery of Art, 2010.

Scholten, Frits, et al. *Art and Migration: Netherlandish Artists on the Move, 1400-1750 = Kunst En Migratie: Nederlandse Kunstenaars Op Drift, 1400-1750*. Brill, 2014.

“The Migrant Artist in Early Modern Time.” *Artists and Migration, 1400-1850: Britain, Europe and Beyond*, by Kathrin Wagner, Cambridge Scholars Publishing, 2017, pp. 2–20.