Kaleb Tsegaye

GEOG788P

**Visualizing and Analyzing Concert Data Patterns with the Songkick and Spotify’s API**

Introduction

Presently, through the advancement of information technology and big data, the general public is able to indulge in analytics that can be used to provide insight into their usage of certain products. For instance, the music streaming platform Spotify is able to give its users year-end reviews of their music-listening behaviors with statistics showing their favorite songs, genre compositions, and even their listening habits by season. Analytical efforts are also given on the artists’ side, with Spotify providing information about the geographic make-up of their listeners and finding the cities their music is the most popular. This information is enough for some independent artists to guide them on their music performance tours (Vice News, 2018). Outside of this these types of analytical services, there is not much to work with. An article from the American Society of Composers, Authors and Publishers, an organization based on supporting American musicians, proposed the utilization of online review platforms, like Yelp, to find suitable venues to tour (Herstand, 2012). This is not bad, since the type of venue can affect the audience that will be likely to attend, but this can be monotonous to find favorable locations, and it does not provide any information on the prevalence of live performances in these areas. To figure this out would require the evaluation of concert data to find feasible and established locations for performing.

Songkick is a well-known platform that harbors a database of over six million concert entries. These entries contain pertinent information such as the exact time and location of a show. With these many entries, it would be beneficial to leverage it to find the best locations that an artist can visit to create an efficient tour path that considers the precedence set by artists before them. Also, we can attempt to find relationships between demographic and socioeconiomic factors and the frequency of shows in areas. The ultimate goal of this paper is to outline my tool for visualizing and examining the patterns of hip-hopconcert performances in the United States, through the use of Songkick’s concert data, Spotify’s artist metadata, and the United States Census Bureau’s census tract data.

An outline of the process taken to create this tool, as well as the platforms used, and their limitations are included. The results of linear regressions done on concert frequency and certain socioeconomic and demographic factors from census tract data will be exhibited and discussed to see how these factors influence show frequencies. A review of the tool current state and future work that can be done will serve as a conclusion.

Related work

Looking for previous applications in mapping concert data or optimizing tour stops gave sparse results. Projects working with concert information were less than likely to be focused on visualizing the information on the platform. Thus, I focused my research on topics that would assist the creation of my project, specifically how I could properly utilize census tract data in regard to concerts.

A paper by scholars from Ithaca College outlines the creation of an application for producing local-music playlists using similar concert information from Ticketfly and Facebook to find artists that have the majority of their performances in a specified area (Akimchuk et al., 2019). Several cities were used to create results for this project. This is supportive of similar goals, specifically providing more attention towards assisting local independent artists and their spatial behaviors. Although it is not as applicable to my work, since it is more focused on artists in small areas, rather than at a regional or national level, and does not have a visualization aspect. This does provide an interesting way to approach retrieving data and algorithms to optimize finding recommended artists.

Making widely accessible user interfaces can be a difficult task, since it is likely to play an important role in keeping users interested in using a product. A paper from the Journal of Visual Languages and Computing documents a platform to make creating web application interfaces simple by allowing users to construct their own based on their needs (Desolda et al., 2017). An example is provided that shows the creation of a map that uses Songkick’s data to visualize an artist’s concert data, by dragging and dropping certain attributes that are desired by the user into the app. This process of allowing the user to mold their own app that is tailored to their needs is significant, as it helps guide thinking how to make a product more accessible and flexible to a user’s needs.

As for finding the demographic and socioeconomic factors that could be examined at a census tract level for an effect on show frequency, a couple studies were used to influence how my regression model was made. One study on residents in the urban English-speaking Canadian cities of Toronto and Vancouver focused on the correlation between enjoyment of certain music genres and class position in society to indicate “highbrow” or “elite” and “lowbrow” or “common” cultures that are present in these cities (Veenstra, 2015). Within the study, it was found that enjoyment of hip-hop was less likely to be associated with individuals that were older or white and it was concluded that hip-hop was considered more accessible to “lowbrow” music listeners. One thing I decided to keep in mind from the study was that listening habits cannot be perfectly predicted, since some of the negative predictors of hip-hop enjoyment also included those with “less than high school” education and those with “middle income,” which are qualities that would match with a lowbrow status. This observation of difficulty in music predictions is related to other studies on the increasing homogeneity of music tastes in current cultural landscapes (García Alvarez et al, 2007). Even though my project focuses on patterns within the United States, the results from the ‘lowbrow’ study should have applications in the US due to the proximity of the cities and the focus on English-speaking urban residents. A paper using music listening data to examine listening conformity between countries showcased that the United States and Canada actually have quite similar listening patterns considering genres and album preferences (Liu et al., 2018). These comparable results mean that it could be fruitful to test on some of their variables.

Data

There were three main sources of data utilized in this project: Songkick for concert data, Spotify for artist metadata, and the US Census Bureau for census tract information. Each source had an Application Programming Interface, or API, to make access of information easier. To get data from Spotify and the Census Bureau, I used the external Python modules “Spotipy” and “census,” respectively, to assist me in making calls to their platforms.

Songkick’s API was possibly the easiest to use, only requiring the correct construction of a URL to get concert information. Their API allows access to concert information, such as the city, latitude and longitude, time and date, and popularity as calculated by Songkick. Only location information and the date and time were needed. Although, information about the artists, like the genres they perform, or gauges of their own popularity are not available. To get this information, I had to use the streaming service Spotify’s API, which provides metadata on millions of artists.

Spotify’s API allows developers access to extensive information on the many artists that use their platform. The main attributes I focused on were the number of followers an artist had, the genres they perform, and their related artists, according to Spotify. Spotify also provides a value for every artist that signifies their current popularity on their platform based on recent song plays, but I preferred follower information since that could be used to gauge an artist’s reach and how established they are rather than a value that could waver.

As mentioned before, a portion of the project was dedicated to finding a correlation between the frequency of shows in a census tract and different socioeconomic factors. I used a linear regression model to find a relationship between median age, the population percentage of black residents, and median household income . This was completed using data from the United States Census Bureau. This was used in conjunction with another package called censusgeocode, which I required to get census tracts from each event’s latitude and longitude information.

For the most part, these sources of data have been fairly reliable. Originally, when I began working with these APIs, I expected there to be more of an issue with rate limiting, but fortunately, this was not an issue. Also, completing the work in Jupyter notebook, which allows you to retain the state of certain variables within different cells, made testing these APIs a simple task. Although, I must mention I did run into trouble with the censusgeocode package, which gave constant errors. This was likely due to the frequency of my calls to the service. Other than this, these sources had been significantly helpful in the creation of my project.

Methods

This project began with the creation of several functions to attain information through Songkick’s API, as well as the beginning of an interface for potential users to access data with. I wrote the function “make\_points,” which retrieves the fifty most recent concert points, and “run\_plot,” which used to plot the resulting points on a map using Matplotlib. These functions were not directly utilized much in the project, but they served as a blueprint for later functions when accessing Songkick. The interface created begins when the “req\_input” function is called, leading to an infinite while loop to allow multiple queries to the Songkick API. This interface allows a user to input an artist and dates to plot their tour path. This interface also allows access to various functions written during the process of this project, which will be described later.

As mentioned before, Songkick does not contain much metadata for artists, so Spotify’s web API was utilized to access their extensive collection of artist information. I wrote a function called “related\_artists,” which takes an artist’s name or Spotify ID and accesses their list of related artists, creates a list of these artists sorted by the number of genres they share with the root artist, and loops through this until ten queries are made to the API. This is to restrict extreme usage of the API to refrain from being limited by Spotify.

The “compile\_points” function returns a geodataframe of the twenty most recent events for each artist in the list returned by the previous function. This allows me to get a decent sample of performance point data. I began to work on functions that would allow me to complete more analytical work. The “basic\_analysis” function, this function currently provides the most frequented venues, cities, and states, makes a choropleth map using the global data for the countries, and provides spatial autocorrelation analysis. This function provided me a decent start for getting simple statistics out of a sample of shows, but I continued to expand my analysis with two functions that would be used to answer more in-depth questions about concert patterns.

I created a function “optimal\_path” that creates and maps a tour path by classifying cities based on the amount of shows they held and plotting a path through points with the most frequent shows. The function takes a list of artists, generates lists of related artists, then retrieves lists of shows for each artist. It then tallies up how many shows were in each city, calculates the mean and standard deviation, then classifies the points in four color categories:

* + - Green - the amount of shows tallied > (mean + 1.5\*standard deviation)
    - Light green - the amount of shows tallied > mean
    - Yellow - the amount of shows tallied > mean – (half of the standard deviation)
    - Red - all unclassified points.

After classifying the points, the optimal path is created. The path begins with every green point. The process of adding to the path is:

* For each unadded point:
  + Check if within a distance between any two points currently in the path. (Light-green: 1.5 \* distance, yellow: 1.25\*distance, red: 1.05\*distance)
  + If so, add to path and restart loop through the unadded points
  + If not, continue through points.

Once every eligible point was added into the path, everything was plotted through Folium, a python package for creating interactive maps in the JavaScript library, LeafletJS. These dipslayed every point, whether or not they were in the path, then plotted the optimal path on top. This provided a user the full scope of the sample along with the locations with the highest event frequencies.

The final task of this project was to examine the relationships between age, race, and income with show frequencies in census tracts through a linear regression model with my function “tract\_reg.” This function takes a list of events, finds their census tracts based on latitude/longitude information, and retrieves data from the Census’s API for each tract. After this, the events in each tract are tallied and a regression is done with the number of events as the dependent variable and non-white population percentage, median age, and median household income as the explanatory variables.

I focused my tests specifically on the genre of hip-hop, based on its widespread popularity currently in America. I compiled lists of lesser-known artists from four regions: East, West, South, and the Midwest. These were the root artists for my optimal\_path and tract\_reg functions. After running these tests with these lists of artists, I compiled the results to look for comparable or significant patterns.

Results

Creating optimal paths for each set of artists produced surprisingly similar plots. Each plot generated some points in their respective regions, but the paths created usually consisted of cities like New York City, New York, Los Angeles, California, and Dallas, Houston, and Austin in Texas. These results made sense, since these are fairly large cities with some having historic and prominent hip-hop scenes. Although, it is interesting to see how prominent these locations are even with these different sets of artists.

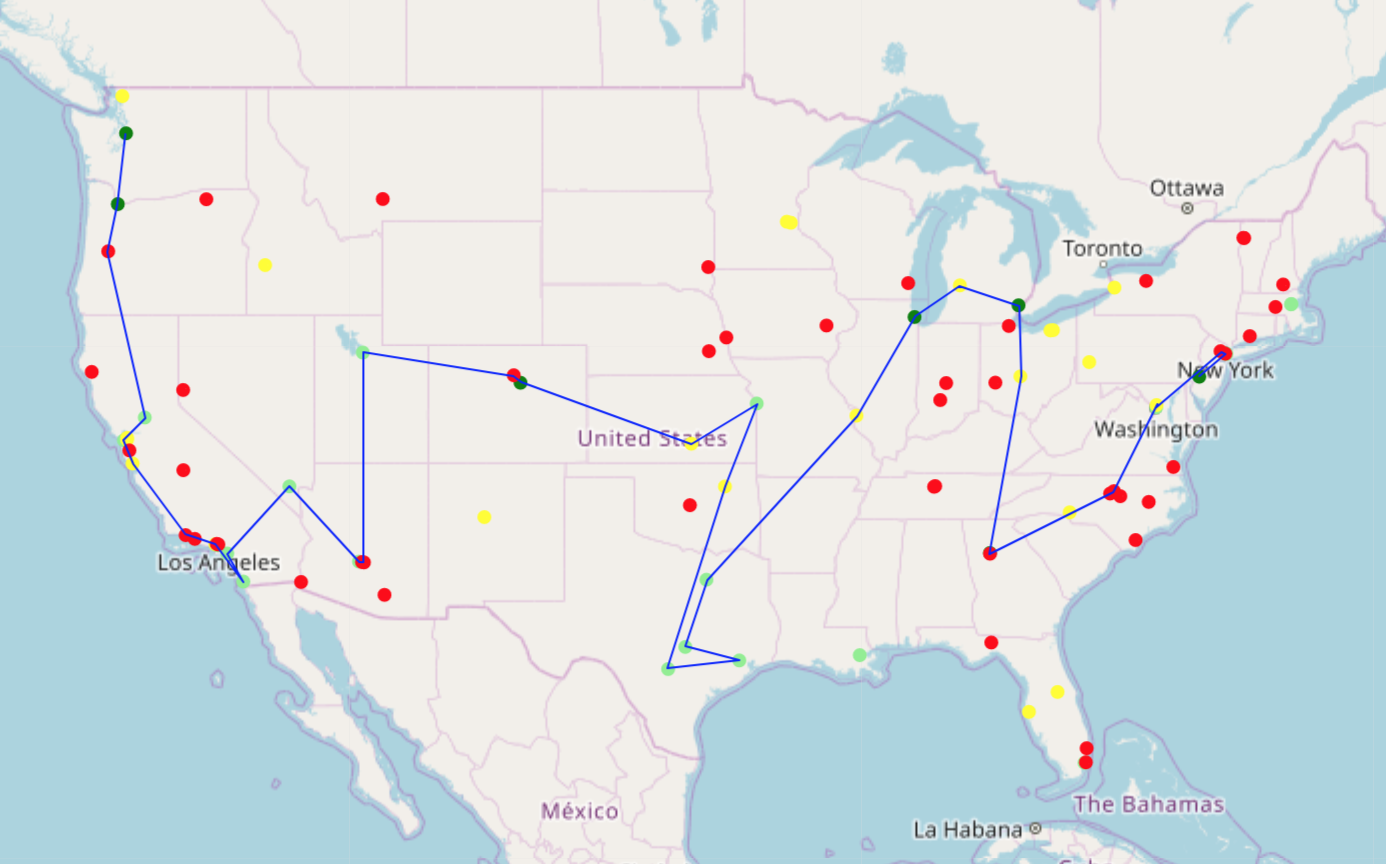


Figure 1) Optimal paths created using artists from the West.

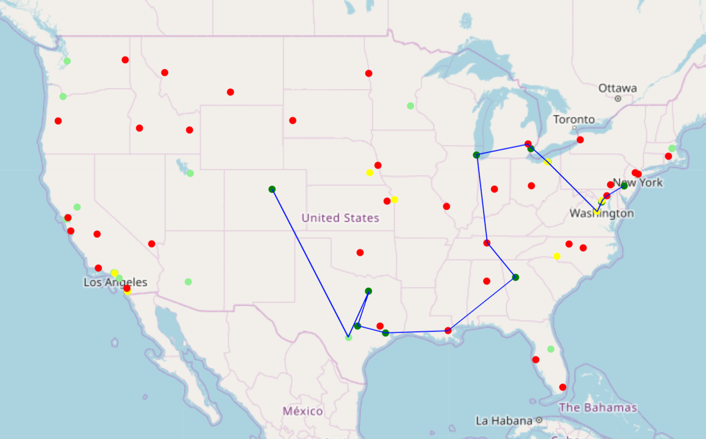


Figure 2) Optimal paths created using artists from the Midwest.

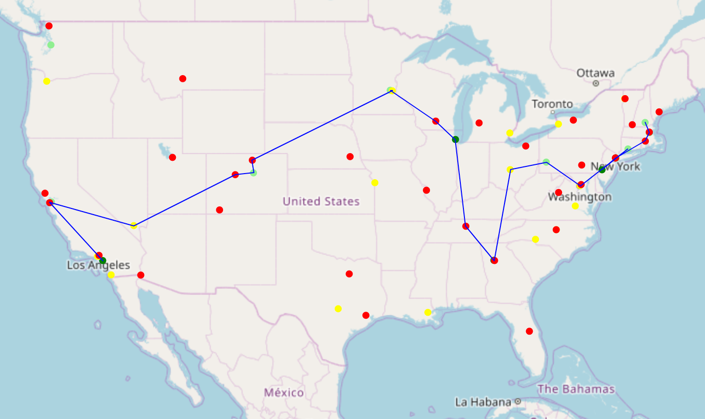


Figure 3) Optimal paths created using artists from the East.

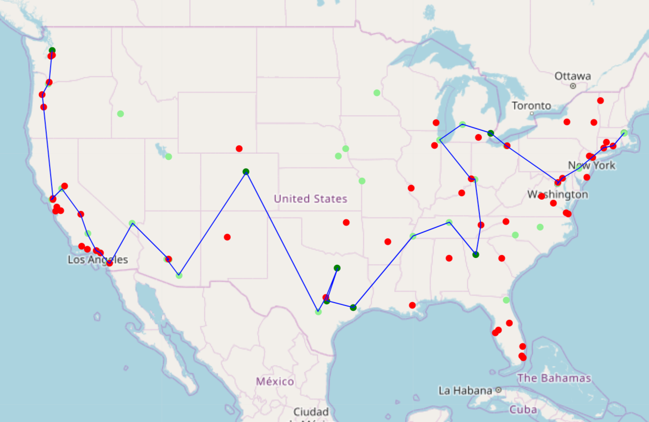


Figure 4) Optimal path created using artists from the South.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **West** | | | | **Midwest** | | | |
|  | **Coeff** | **Std Error** | **t** | **p-value** | **Coeff** | **Std Error** | **t** | **p-value** |
| Black Pop % | -0.2077 | 0.399 | -0.521 | 0.604 | 0.0172 | 0.507 | 0.507 | 0.973 |
| Median age | 0.0325 | 0.005 | 6.890 | 0.000 | 0.0272 | 0.007 | 4.128 | 0.000 |
| Median household income | 3.488e-06 | 1.94e-06 | 1.797 | 0.075 | 5.296e-06 | 3.2e-06 | 1.657 | 0.104 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **East** | | | | **South** | | | |
|  | **Coeff** | **Std Error** | **t** | **p-value** | **Coeff** | **Std Error** | **t** | **p-value** |
| Black Pop % | -0.2158 | 0.453 | -0.477 | 0.635 | 0.4509 | 0.337 | 1.340 | 0.183 |
| Median age | 0.0416 | 0.004 | 10.893 | 0.000 | 0.0211 | 0.005 | 4.661 | 0.000 |
| Median household income | -3.17e-10 | 1.02e-09 | -0.311 | 0.757 | 5.671e-06 | 2.15e-06 | 2.640 | 0.010 |

Tables 1, 2) Displaying the results of the linear regression on groups of artists from four regions.

Using the same lists of artists, I ran a regression to see the relationships between non-white population percentage, median age, and median household income and show frequency, and my results were barely conclusive (Tables 1,2). Median age was shown to be positively and significantly related to show count, but its effect was quite minuscule. Median household income was also seen to be significantly related within the West and South tests, with a positive coefficient. After running these tests, I then decided to make some adjustments. I considered that even though younger people were more likely to listen to hip-hop, they were relatively less likely to be financially established and therefore attend a show than compared to older people. So, age would be positively influential until a point where the likelihood of listening to hip-hop decreases. I also considered that people with a low income would be less likely to attend a show compared to those with a middle-class background due to their financial limitations, and those with high incomes would be less likely to indulge in the genre. To adjust for these considerations, I squared both the median age and median household income variables, since this would accommodate for their potential parabolic behavior. I also applied the log function to the black population percentage to compare with my current results.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **West** | | | | **Midwest** | | | |
|  | **Coeff** | **Std Error** | **t** | **p-value** | **Coeff** | **Std Error** | **t** | **p-value** |
| log(Black Pop %) | -0.1551 | 0.059 | -2.619 | 0.010 | -0.1205 | 0.076 | -1.594 | 0.116 |
| (Median age)^2 | 0.0007 | 0.000 | 5.140 | 0.000 | 0.0010 | 0.000 | 5.664 | 0.000 |
| (Median household income)^2 | 2.235e-11 | 1.45e-11 | 1.545 | 0.125 | -1.624e-11 | 1.34e-11 | -1.213 | 0.230 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **East** | | | | **South** | | | |
|  | **Coeff** | **Std Error** | **t** | **p-value** | **Coeff** | **Std Error** | **t** | **p-value** |
| log(Black Pop %) | -0.1816 | 0.078 | -2.336 | 0.023 | -0.1400 | 0.061 | -2.291 | 0.025 |
| (Median age)^2 | 0.0007 | 0.000 | 4.554 | 0.000 | 0.0005 | 0.000 | 3.959 | 0.000 |
| (Median household income)^2 | -3.486e-12 | 1.96e-11 | -0.178 | 0.859 | 2.361e-11 | 1.28e-11 | 1.847 | 0.068 |

With these adjusted variables, there was a visible negative relationship between the black population percentage in a census tract and the amount of shows in the census tract. Age continues to show a miniscule, but statistically significant positive relationship and median household income still does not have much of a significant effect on the amount of shows in a census tract. My results went against the predictions I had for the relationship with these variables. It must also be mentioned that these factors don’t necessarily reflect the demographics of the audiences of the venues, since these people can come from other areas, but in the absence of this information, I used this as an alternative for predicting the best venues.

Discussion and Conclusion

* Geodemographics, ml algos, travelling salesman – try to implement costs, improve path

The resulting culmination of my work ended up with a good basis for concert visualization and analysis. Through this platform, a user would be able to easily plot a single artist’s shows, multiple artists’ tours, and even find the most frequent locations artists visit on tour with a plot of a path through these points. Patterns in socioeconomic and demographic factors can be used in linear regression models to detect patterns at the census tract level. Even though, I had completed these tasks, I do not believe my project is near the point of completion.

Some improvements can be implemented in the presentation of my maps. Presently, the maps are quite simplistic and while they do accurately represent what they are tasked to show, they could definitely use more work to provide users more important information. Plotting clusters and adding a legend to the folium maps, as well as making the static matplotlib maps more descriptive are some oversights I should have considered. It would be beneficial to model my future work based on issues detailed in a paper on Research Challenges in Geovisualization. . Some concepts outlined regarding visual computation integration specifically focus on data mining, include incorporating temporal aspects of data further into the visualization and taking into account how “human inference processes” can affect how visualizations are perceived (MacEachren and Kraak, 2001). This in consideration with pertinent qualities for automated maps could be helpful (Pillay et al, 2019).

As mentioned earlier, functions that complete basic analysis of concert data were created that provided autocorrelation plots. This application could be expanded to add a time dimension to examine how concentration of shows in states and countries change over time. This can be applied through directional LISA plots that visualize how clusters move from one plot to another (Rey, 2014). This could be useful for displaying how hot spots and cold spots for concerts of a certain genre or set of artists migrate. This can be combined with aggregated trajectory plotting for tour paths, which would allow the ability to share patterns from large sets of data based on space-time density (Andrienko et al., 2010, Laube et al., 2007).

One of the main issues with my optimal tour path function is that while it does include points of higher frequency and manages to fit in points of interest based on distance to the path, the final path can become unnecessarily lengthy and redundant. For example, the path can pass a city it eventually adds later, causing the path to have to loop back to get to the point. Optimization is required to deal with this, and the utilization of graph networks could assist this task by making it easier to find shortest paths with a multitude of edges from point to point (Batty, 2003). The creation of these paths should also be able to take into account the frequency of origin and destination cities for each point, and the concept of distance decay, specifically in the likelihood of artists travelling large distances between shows (Farmer and Oshan, 2017). These factors can provide more realistic paths, especially if a path suggests travelling a large distance to a region that only has one applicable city or venue, even if it has a high tally of shows. Also, regarding realistic paths is an application of the Travelling Salesman Problem tailored towards transportation costs, which would be beneficial in the case that it is possible to find costs of travelling from city to city while touring on a state or national level (Maggioni et al, 2014).

References

Akimchuk, D., Clerico, T. and Turnbull, D., (2019), Evaluating Recommender System Algorithms

for Generating Local Music Playlists. arXiv:1907.08687.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S., Jern, M., Kraak,

M., Schumann, H. & Tominski, C. (2010). Space, Time and Visual Analytics. International

Journal of Geographical Information Science. 24. 1577-1600.

10.1080/13658816.2010.508043.

Batty, M. (2003). Network Geography: Relations, Interactions, Scaling and Spatial Processes in

GIS. https://www.researchgate.net/publication/32884874\_Network\_Geography\_Relat

ions\_Interactions\_Scaling\_and\_Spatial\_Processes\_in\_GIS

Desolda, G., Ardito, C., Costabile, M.F., Matera, M. (2017), End-user composition of interactive

applications through actionable UI components, Journal of Visual Languages &

Computing, Volume 42, 2017, Pages 46-59, https://doi.org/10.1016/j.jvlc.2017.08.004.

Farmer, C. and Oshan, T. (2017). Spatial interaction. The Geographic Information Science &

Technology Body of Knowledge (4th Quarter 2017 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2017.4.5

García Alvarez, E., Katz-Gerro, T., & López Sintas, J. (2007). Deconstructing Cultural

Omnivorousness 1982-2002: Heterology in Americans' Musical Preferences. Social Forces 86(2), 417-443. https://www.muse.jhu.edu/article/231541.

Herstand, A. (2012), Booking Your Own Tour: A How-To Guide.

<https://www.ascap.com/help/career-development/booking-your-own-tour-a-how-to-guide>

Laube, P., Dennis, T., Forer, P., & Walker, M., (2007), Movement beyond the snapshot –

Dynamic analysis of geospatial lifelines, Computers, Environment and Urban Systems, Volume 31, Issue 5, 2007, Pages 481-501, https://doi.org/10.1016/j.compenvurbsys.2007.08.002.

Liu, M., Hu, X., & Schedl, M. (2018), The relation of culture, socio-economics, and friendship to

music preferences: A large-scale, cross-country study. PloS one, 13(12), e0208186. doi:10.1371/journal.pone.0208186

MacEachren, A. & Kraak, M. (2001), Research Challenges in Geovisualization, Cartography and

Geographic Information Science, 28:1, 3-12, DOI: 10.1559/152304001782173970

Maggioni, F., Perboli, G., & Tadei, R. (2014), The Multi-path Traveling Salesman Problem with

Stochastic Travel Costs: Building Realistic Instances for City Logistics Applications,

Transportation Research Procedia, Volume 3, 2014, Pages 528-536,

https://doi.org/10.1016/j.trpro.2014.10.001.

Pillay, L., Schaab, G., Coetzee, S., & Rautenbach, V. (2019), A comprehensive workflow for

automating thematic map geovisualization from univariate big geospatial point data. In Proceedings of the ICA (Vol. 2).

Rey S.J. (2014), Spatial Dynamics and Space-Time Data Analysis. In: Fischer M., Nijkamp P. (eds)

Handbook of Regional Science. Springer, Berlin, Heidelberg

Veenstra, G. (2015), Class Position and Musical Tastes: A Sing‐Off between the Cultural

Omnivorism and Bourdieusian Homology Frameworks. Canadian Review of Sociology/Revue canadienne de sociologie, 52: 134-159. doi:[10.1111/cars.12068](https://doi.org/10.1111/cars.12068)

Vice News. (2018), How Grammy Nominee Brent Faiyaz Built His Music Career Off Streaming.

Retrieved from https://www.youtube.com/watch?v=JTaeDLrj7q8