Title: Final Project Prospectus
Notice: Dr. Bryan Runck
Author: Michael Felzan
Date: February 20, 2012

Project Repository:
https://github.com/fezfelzan/GIS5572.git (navigate to "5572 FINAL PROJ REPO")

Abstract

This project explores the variation in styles of radio-played music across the US in semi-real-time. The website "radio.garden" will be used for the main source of radio data, and audio information from various radio station's live streams will be the source data for "popular" local music. Clips from each radio station's stream will be collected and programmatically routed into acrcloud.com's API to return song name information, which will be further routed into a music library database's API to return aggregated song genre information. Understanding the styles of music that receive the most airplay in different areas of the US may allow music-oriented marketing decisions may be made on a daily, monthly, or yearly scale.

Problem Statement

The objective of this project is to write a Python program in the Esri ArcGIS environment which programmatically records short audio clips from a series of self-selected radio stations across the US, routes the recordings to a music song identifier API, in order to further route the song names into a music library database – ultimately returning data which describes the highest frequency genres of songs across the US. This project assumes that the songs which receive airplay by radio stations may be used as a proxy for determining the spatial variation in music preferences. The end product could theoretically be used for marketing purposes, as information about localized music tastes could inform tour routes, electronic marketing campaigns, and decisions made by DIY musicians. Though the main intent of this project is to develop an ETL methodology which uses web-scraping and API requests, in order to develop a novel database of general music preference trends across geographic space.

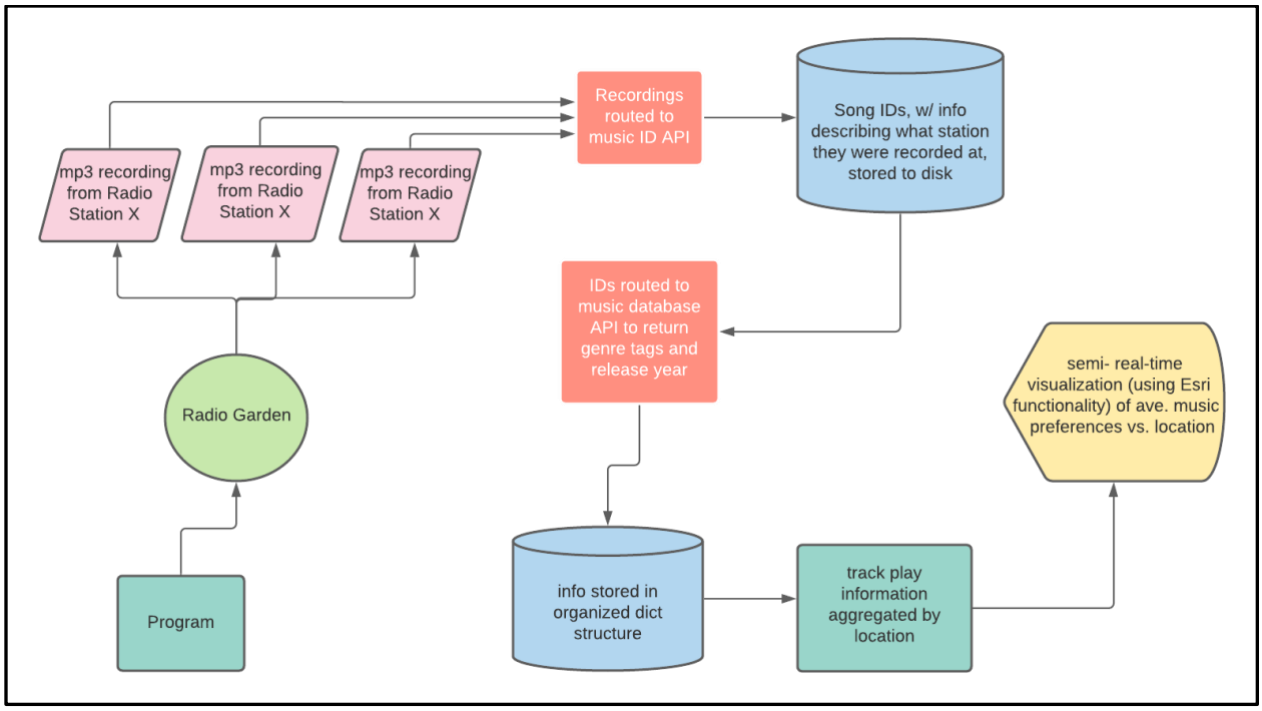

Figure 1. Flow Map depicting process of programmatically tracking types of songs receiving most airplay, on average, in different areas in the US

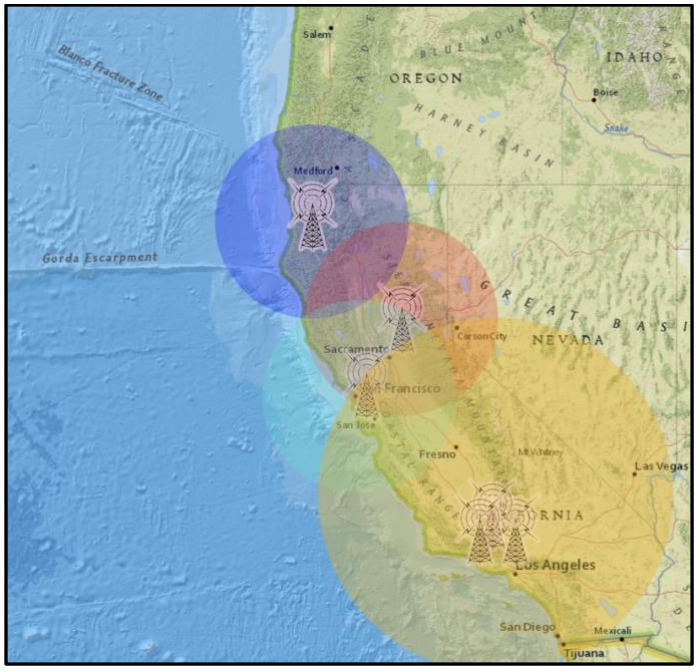

Figure 2. Illustration/ sketch visualizing overlapping broadcast reaches on the Western US Coast. Each buffer will contain attribute information on frequency of each song genre played in that area; areas between overlapping buffers will be averaged using a weighted formula which considers the distance from each station, and the number of listeners tuned in to each.

Table 1. Deconstructed Problem Statement

#	Requirement	Defined As	Spatial Data	Attribute Data	Dataset	Preparation
1	Broadcast "width" for each radio station	(will be different for each station) X meters, based on station reach	Circular buffer around point locations		(Will be programmatically created from mining http://radio.garden/)	Selecting popular/"influencing" radio stations, gathering location info & signifier ID for each station
2	Songs played on radio	~5 second mp3 (which yield result from music track name identifier API)		Digital Audio	Will be collected via http://radio.garden/	Building program that records 5 second mp3 clip from stations on http://radio.garden/
3	Radio stations	Manually/ programmatically inputted locations with ID signifiers	Point Locations	Station ID; City info		
4	Music preference	Top (on ave.) music genre tags & release year		Genre tags, release year	Wikipedia database? Spotify database???	Sending track IDs to database, receiving list of genre tags/release year, aggregating info, averaging preferences across intersecting broadcast buffers

Input Data

The primary data that will be used in this project will be sourced from the RadioGarden website. Using a combination of developer tools and Python functions, the geographic coordinates and unique ID's of each relevant radio station will be captured into a Python dictionary structure. The geographic coordinates of each radio station, along with the broadcast reach and number of listeners for each station, will be a proxy for a geographic location's musical preference. Other data integral to this project will be short .mp3 audio clips captured from every station at consistent time intervals. This data will be used to determine the nature of music most frequently played by stations in certain areas. A shapefile containing US borders will also be necessary for this project.

Table 2. Input Data Descriptions

#	Title	Purpose in Analysis	Link to Source
1	Radio station locations & RadioGarden station IDs	Mapping musical preference against geographic location	http://radio.garden/
2	Mp3 clips from radio stations across US	Pieces of audio from full songs played on the radio – to determine what (types of) songs are being played across the US	http://radio.garden/
3	Shapefile of US	Geographic reference for radio station locations	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

Methods

The methodology for this project is outlined in Figure 1. The first step of this process is to select a series of radio stations which represent major cities in the U.S., and have substantial "influence" over the area, characterized mainly by number of listeners. Once the stations have been selected, their unique "ID's" or URL paths on http://radio.garden/ will be inputted to a Python dictionary structure. A method will be developed which iterates over all of these radio streams at specific times of the day, captures mp3 clips from each station, saves these clips to disk memory to further route them to a song identifier program's (https://www.acrcloud.com/) API, in order to return the song titles for each of these recordings (a method will be developed which accounts for the possibility that a radio station DJ is talking during those periods, or if an advertisement is playing). Once a comprehensive list of song titles has been generated for each radio station over a defined time-frame, the song titles will be routed to another music library database's API (possibly Wikipedia's) to return genre tags to the previously-built dictionary structure. The genre tags will be aggregated, so that ultra-specific genre names like "down-tempo" may be grouped together with "trip hop," while overly vague genre titles such as "R&B" or "electronic" may be disregarded. Once genre frequency statistics have been generated for each broadcast area, a method will be created (described in Figure 2.) which uses weighted averaging to determine the statistics of areas that share overlapping broadcast reaches. Using a collection of map algebra techniques, an interactive web map surface will be created which visualizes the results of this program in close to real-time.

Results

TBD

Results Verification

TBD

Discussion and Conclusion

TBD

References

https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

http://radio.garden/

https://www.acrcloud.com/

https://en.wikipedia.org/