Summer 7-13-2017

Incorporating Census Data into a Geospatial Student Database

Edwin Agbenyega
University of New Mexico

Edwin Agbenyega

Candidate

Electrical and Computer Engineering

Department

This thesis is approved, and it is acceptable in quality and form for publication:
Approved by the Thesis Committee:

Prof. Gregory L. Heileman, Chairperson

Prof. Don R. Hush

Dr. Heather Mechler

Incorporating Census Data into a Geospatial Student Database

by

Edwin Agbenyega

B.Sc., Computer Engineering, Chungnam National University, 2014

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Engineering

The University of New Mexico

Albuquerque, New Mexico

July, 2017

Dedication

To my parents, Emmanuel and Happy, for their support and encouragement throughout the years.

Acknowledgments

I would like to thank my advisor, Professor Greg Heileman, for his support and guidance over the past two and a half years of graduate school.

I would also like to thank my other committee members, Heather Mechler and Don Hush. Especially Heather, for sacrificing several hours to proofread my thesis and for providing very constructive feedback.

To my friends and family, I am very grateful for your continued support and encouragement as I worked on my thesis. This would not have been possible without you.

Incorporating Census Data into a Geospatial Student Database

by

Edwin Agbenyega

B.Sc., Computer Engineering, Chungnam National University, 2014

M.S., Computer Engineering, University of New Mexico, 2017

Abstract

The University of New Mexico(UNM) stores data on students, faculty, and staff at the University. The data is used to generate reports and fill surveys for several local, statewide and nationwide reporting entities. The reports convey statistical and analytical information such as the graduation rates, retention, performance, ethnicity, age, and gender of students. Furthermore, the Institute of Design and Innovation (IDI), and the Office of Institutional Analytics (OIA) at UNM use the data provided for various predictive studies aimed at improving student outcomes.

This thesis proposes geospatial data as an additional layer of information for the data repository. The paper runs through the general steps involved in setting up a geospatial database using PostgreSQL and geospatial extensions including PostGIS, Tiger Geocoder, and Address Standardizer. With geospatial functionality incorporated into the data repository, the university can know how far students live, which amenities are in proximity to students, and other geospatial features which describe students' journeys through college.

To demonstrate how the university could exploit geospatial functionality a dataset of UNM students is spatially joined to socioeconomic data from the United States' Census Bureau. Various student related geospatial queries are shown, as well as, how to set up a geospatial database.

Contents

Contents

List of Figures

List of Figures

List of Tables

Acronyms

2D two dimensional.

3D three dimensional.

4D four dimensional.

ACT American College Testing.

CGPA Cumulative Grade Point Average.

ERD Entity Relationship Diagram.

FPL Federal Poverty Level.

GIS Geographic Information System.

GPA Grade Point Average.

GRAPI Gross Rent As a Percentage of Household Income.

HEI higher education institution.

HSGPA high school Grade Point Average.

NoSQL Not only SQL.

Acronyms

OGC Open Geospatial Consortium.

SAT Standard Aptitude Test.

SQL Structured Query Language.

TIGER Topologically Integrated Geographic Encoding and Referencing system.

U.S United States.

UNM the University of New Mexico.

Glossary

geocoding Converting addresses into latitude and longitude values that could be mapped to a point on the earth's surface.

geospatial data Data that provides information about geographic locations on the earth.

geospatial database A database that stores geospatial data and supports geospatial queries.

geospatial queries A type of query that is supported by geospatial databases.

relational database A SQL based database that stores data in columns and rows.

Student Data Mart A data repository for student data at the University of New Mexico.

time-to-degree The amount of time a student takes to graduate, or the number of terms a student is enrolled and registered, from matriculation to graduation.

Chapter 1

Introduction

Over the past few years, geospatial technology has grown to encompass a wide variety of tools and disciplines that deal with geographic mapping and analysis of the Earth and human societies. These include remote sensing imagery tools, Geographic Information System (GIS) software, Global Positioning Systems (GPS) satellites and receivers, and internet mapping applications such as Google Earth, Google Maps APIs, and OpenMap [1]. Analysis of the data generated by these tools provides information on various socioeconomic and environmental issues in different parts of the world.

This thesis examines the incorporation of such analysis into higher education by evaluating how socioeconomic and geographic data could be extracted from the United States (U.S) Census Bureau and other data sources. It also considers how the performance of students from a given geographic region could help advise incoming students. The dataset used for the study is based on student data derived from the University of New Mexico (UNM)'s data repository. Statistical information from the study is presented to illustrate how various geospatial queries on census data, when integrated with student data, can provide substantial information on a student's

performance and give insight into ways to improve student outcomes. In this thesis, student outcomes and student performance may be used interchangeably to mean a measure of student success regarding grades, retention, and time to graduation, also known as the time-to-degree.

A major focus is how geospatial data is stored using a PostgreSQL database with extensions that provide geospatial functionality, as well as, how the information is processed with various software packages and presented to the end user. From a software developer's perspective, there are two principal aspects worth considering to provide the end user with the necessary data: back-end processing and front-end presentation. Back-end processing includes how the data is processed, stored, queried, or analyzed. Front-end presentation deals with visually presenting the information to the user via a front-end interface. While the analysis involves dashboards of charts and maps, the study focuses more on processing and querying the information. This thesis also demonstrates the steps involved in analyzing student data based on socioeconomic information on the geographic locations of students, using geospatial joins and queries.

Various higher education institutions (HEIs) around the country increasingly study student outcomes [2]. In particular, which parameters are predictive of student outcomes and what measures could be taken to improve results. This is due to a demand for increased accountability, declining state allocations, more diversity of student populations, and the overall expansion of higher education in the last half century [3, 2]. Various data analysis tools are used to this end, and several algorithms are adopted to analyze curricular information based on designated factors such as ethnicity, gender, age, and location. This thesis takes a similar approach, with an emphasis on geospatial information and how it can be used to profile a student. By mapping out the location of students before, during and after their time at the university, informed decisions can be made to address various problems students face.

For example, exploring the most desirable jobs within specific geographic regions helps to advise students on their post graduation prospects. Another use case is to help students plan out their class schedules based on their proximity to the university and access to transportation. While this thesis does not delve into all the problems geospatial information addresses, it answers some fundamental questions that serve as a starting point and explores best practices for analyzing geospatial data.

The primary focus of the thesis involves setting up a geospatial database for an educational institution. However, it is also essential to explain why a geospatial database is necessary. Thus, subsequent chapters consider the importance of geospatial data by highlighting some queries that address geospatial questions. The study explores geospatial information of an area and how it affects the students from the area and examines how the performance and choices of students within an area describe and influence the geospatial region. Thus, the geospatial queries addressed by the paper may fall within the following inquiries. a) Do the attributes of students from a particular region provide information about the region? b) Do socioeconomic data based on students' geographic origins relate to student outcomes and choices? c) What geospatial information can be gathered about students and academic institutions?

The first question addresses potential issues or traits within a geographic location, based on choices and outcomes of students living in the region. The data considers whether there is a balance of males and females from a particular geographic region, whether some areas do not produce students in specific academic disciplines, and whether there are concerns (such as retention problems and low Grade Point Average (GPA)) that need to be addressed regarding students from a particular geographic location. The second question encompasses issues such as, which major a student is likely to pick, most preferred colleges, and the tendency of students to switch or drop majors, based on socioeconomic information about a student's geo-

graphic origin. Finally, the third question explores relevant geospatial information about students and institutions outside of census data. This includes the proximity of students to libraries, concentration of students in urban and rural areas, the infrastructure of geographic regions around academic institutions, and potential benefits or disadvantages of living in a geographic area.

The second chapter delves into the literature concerning outcomes in higher education and geospatial data, possible applications of results from the study, and the data sources that feed the test applications used in the study. The third chapter introduces the methods used to extract, format and query geospatial data for the study and some of the database tools employed for this purpose.

Based on selected parameters, the fourth chapter looks at the methods used for analysis and how the data is illustrated via maps, graphs, and tables in a user-friendly manner. The thesis concludes with a preview of possible integrations of the work into future applications, as well as, a summary of the work. The appendices consist of source code, technical screenshots of the implementation process, and a list of references.

Chapter 2

Background

2.1 Motivation

The University of New Mexico(UNM) boasts a large and diverse student body. In 2016, out of a total enrollment of 34,674 students, 42.1% were Hispanic, 36.7% white, 5.2% American Indian, 4.9% Foreign, 3.7% Asian, 2.3% African American/Black, 1.8% had no known ethnicity, and 3.2% had more than one race [4]. With such a large and diverse student body, it is essential that the university is aware of the needs of various groups. Students from different backgrounds and students with special needs may require different services or communal groups within the university to better deal with their particular needs. For example, first-generation students are known to have problems navigating through college [5]. Setting up groups for first-generation students from similar geographic backgrounds could ease them into college life. Thus, understanding a student's socioeconomic background plays a significant role in the university's ability to provide services and advise students to improve student outcomes.

At UNM, prior work has been done to predict student outcomes based on some

variables including students' ethnicity, gender, retention rates, courseload and other curricular factors [6]. While this paper does not engage in predictive analysis, it introduces additional variables related to a student's socioeconomic background that could be incorporated into predictive studies.

2.2 Data Sources

2.2.1 University of New Mexico Student Data Mart

UNM has a number of database systems responsible for the storage of student information. Among these database systems is the Student Data Mart, which is the central data store for student related information. Some tables from the data mart were queried to provide relevant student information for the purpose of the study. Some of the data sampled from the Student Data Mart include geospatial information such as students' residential and mailing addresses, and students' high school addresses, as well as other information relevant to evaluating student outcomes. The information includes students' gender, ethnicity, age, grades, time to degree, and other pertinent data.

For illustration purposes, the dataset chosen from the student database consists of undergraduate students living in New Mexico who graduated between fall 2007 and fall 2016. Appendix A provides more information about the dataset and fields used.

2.2.2 U.S Census Bureau (FactFinder)

Geospatial data can illustrate how external factors outside of a school setting may play a part in student outcomes or provide further information that can be

considered when advising a student. The U.S Census Bureau provides data about the socioeconomic status of various census tracts. Census tracts are defined by the U.S Census Bureau as small, relatively permanent statistical subdivisions of a county. By relating this census data to students via the census tracts they live in, we can gain better insight into how students' backgrounds may affect their performance in an institution.

For the study, information was derived from the U.S Census Bureau. This comprised of data on all census tracts within the state of New Mexico based on five main data profiles provided on the website. These data profiles were categorizations of census information consisting of selected social, economic and housing characteristics in the U.S, and the American Community Survey (ACS) demographic and housing estimates.

To facilitate easy access to census data, the Census Bureau provides a web application called the *FactFinder* which has advanced search and filtering options for the Census data repository. The FactFinder web application provided socioeconomic information about various census tracts. However, to associate the data with the students at the university, there was a need for a means to connect both datasets. Census tract shapefiles that contained geospatial information necessary for joining the student data with the Census Bureau's Socioeconomic data were also available on the Census website.

2.2.3 Arbitrary Shapefiles and Geospatial Data

Besides data from the U.S Census Bureau, it was essential to find out how other geospatial data related to student data. For this reason, other arbitrary geospatial datasets were explored. Some sources such as *data.gov* and *Tiger* data provided relevant information to the study. Examples of these data include shapefiles and

geospatial data on federal highways, urban and rural areas, zip code boundaries, and landmarks.

2.3 Test GeoSpatial Database

While the data sources above are the primary data sources for the study, it became apparent that a local geospatial database was needed to ease querying of the data. As such, a smaller geospatial database was created with appropriately formatted datasets derived from the previously mentioned data sources. The database is referred to in the rest of the paper as the test database.

Data extracted from the Student Data Mart, together with the Census Bureau's socioeconomic profiles, the census tracts shapefiles and other resulting files from calculations using Python were imported into a PostgreSQL database. Also, to enable geospatial joins and queries, the PostGIS extension was installed into the database. The test database, therefore, provided a central data store for the geospatial queries presented in the study. The database's extensions, PostGIS, Tiger and Address Standardizer, provided extra functionality for standardizing addresses, geocoding, geospatial querying, joining, and indexing.

The Entity Relationship Diagram (ERD) in Figure 2.1 shows a simplified version of the database structure. In the diagram, there are four main database tables with a fifth join table that links student data to their addresses. The main idea is to spatially join student data to census socioeconomic data via the intermediary geospatial database tables; namely, the *census shapefile* and *student addresses* tables. One notable attribute of the geospatial tables is a `geom` field that utilizes the geometry formats available through PostGIS. The geometry format is not inherently available as a data structure and is provided with the PostGIS extension to allow for geospatial queries and joins [7].

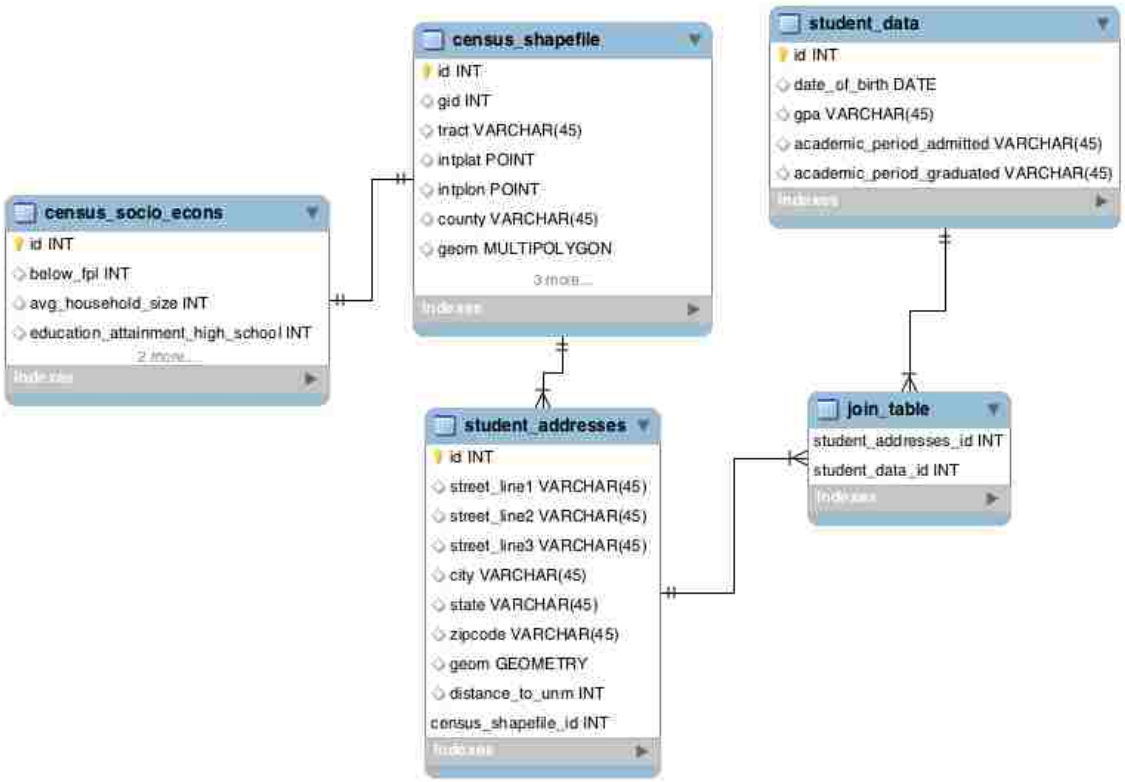


Figure 2.1: An Entity Relationship Diagram for the Test Database

2.4 Project Summary

The creation of a geospatial student database with census data involved four main steps as shown in Figure 2.2. The data was first collected from various sources including the Student Data Mart, the census bureau and *data.gov*. Since the data were from different sources and in different formats, the next step was to format the data and prepare it for the database. The data were then consolidated into the test database. Once the data were available in a central location, queries and analysis could be carried out. The various steps are elaborated in the following chapters, together with the software tools employed for querying and analyzing the data.

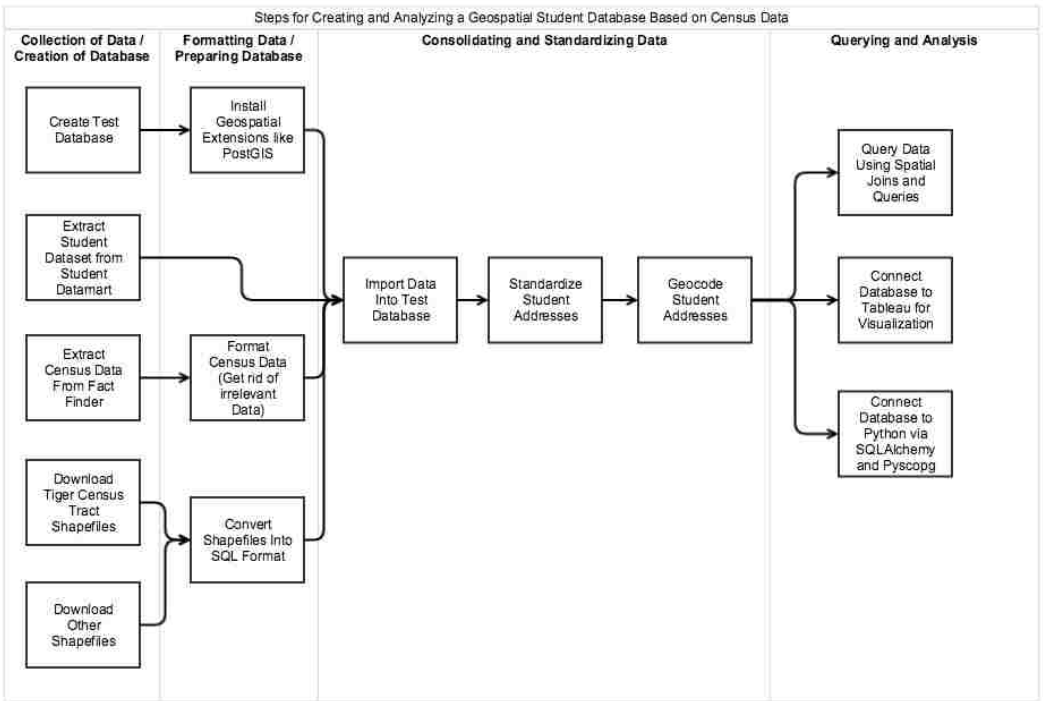


Figure 2.2: Flow Diagram for Process Involved in the Study

Chapter 3

Geospatial Data Processing

To appreciate the need for a geospatial database, it is essential to understand what it is and how it differs from other databases. Geospatial databases are databases that contain geospatial data and can handle geospatial queries. *Geospatial data* refer to data about some point or space on the earth while *geospatial queries* refer to queries that extract relevant information from geospatial data. For the purpose of restricting the scope of data, queries, and functions to geographic locations on the earth, the term *geospatial* replaces the term *spatial*, which is used more generally to relate to any given space. It should be noted that while geospatial databases are mainly used in association with the earth, advances in neuroscience and architecture have increasingly lead to incorporating spatial data structures when storing data about the brain, structures, and other complex three dimensional (3D) objects [8, 9].

Geospatial databases are not inherently different from other databases. Rather, they may either be extensions of regular databases with added geospatial functionality or specialized databases for storing and querying geospatial data. Currently, the two main types of database systems are *Not Only SQL* (NoSQL) databases and relational databases [10]. NoSQL and relational databases handle geospatial data

differently but have some similarities [10]. Details of how NoSQL and relational databases handle geospatial functionality is provided in Appendix B. PostgreSQL is a relational database that was chosen for the test database because of its extensive geospatial library and reputation as a stable and reliable database.

The main differences between geospatial and non-spatial databases lie in the data types, the additional queries, and the types of indexes used in the database [11]. First of all, most databases store two dimensional (2D) data while geospatial databases store both 2D and 3D data. Using arrays as an analogy, a list of student names could be seen as a one-dimensional array, while a 2D array could be a tabular display spanning multiple columns of data consisting of student names, grades, ages and ethnicities. 3D data include entries that convey information about a given 3D space or a point within it. A student's address may be stored as a string in a 2D database but stored as geometric coordinates within a 3D database. While some geospatial databases may not necessarily have 3D geospatial data types, a single geospatial column entry may comprise of several thousands of latitude and longitude coordinate values or a single latitude and longitude coordinate that relates to a point on the earth. Conventional data types are unable to store such information properly.

Geospatial data is stored in a geospatial table using geospatial data structures, which are mostly derived from extensions or libraries added to the database. Since geospatial data is stored using 3D data structures, it follows that binary tree indexes designed for 2D data structures are not efficient for geospatial data. Thus, most databases use geospatial indexes that work differently when indexing geospatial database tables. Relational geospatial databases typically use R-Tree indexing or variants known as R* and R^+ or kd-trees [11]. NoSQL databases also have special geospatial indexes for geospatial entries.

To map census data to student data, a PostgreSQL database was created to store selected student data, socioeconomic Census Bureau data and geospatial data from

the required shapefiles. The student dataset consists of students who graduated between fall 2007 and fall 2016 and lived in New Mexico at the time. The Address Standardizer and Tiger Geocoder extensions were installed to handle standardizing addresses and geocoding the student and high school addresses from the Student Data Mart. PostGIS was installed as an extension to handle geospatial storage, queries, and joins. For further computational and analytical purposes, Python's Pandas and Statsmodels packages were used. Since data from different sources and formats were used, it was essential to standardize the data. This chapter explores the steps taken to format the data and how PostgreSQl and PostGIS are used to store, query and analyze the data.

3.1 Geospatial Data Structures in PostGIS

Geospatial data is usually thought of as map data, which is mostly true. Most information related to geographic information can be illustrated via maps. It follows that maps form a large part of geospatial analysis. How then is a map stored within a database of rows and columns when it contains arbitrary lines, shapes, and colors? There are several formats in which geographic data is stored. A majority of these formats are based on standards developed by the Open Geospatial Consortium (OGC) through a consensus process. Two major categories of Geographic Information System (GIS) file formats are the *raster* and *vector* formats. Raster data is made up of pixels or grid cells stacked in rows and columns while vector data consists of points and paths [7].

PostGIS provides several data structures or GIS objects based on the "Simple Features" defined by the OGC. PostGIS also extends the standard with support for 3D and four dimensional (4D) coordinates [7]. Four main GIS objects are used in the test geospatial database as listed below:

- POINT (latitude and longitude pair coordinates of student addresses, point landmarks, public facilities)
- MULTILINESTRING (roads and highways)
- POLYGON (spatial joins and specific geographic area information)
- MULTIPOLYGON (area landmarks, census tracts, counties, school districts)

The full list of GIS objects and functions supported by PostGIS can be found in the PostGIS manual [7]. To carry out geospatial operations and joins, a `geom` column needs to be included in any geospatial database table. The field type should be one of the geospatial data structures listed above or in the PostGIS manual [7].

3.2 Geocoding Solutions

Geocoding involves converting addresses into latitude and longitude values that could be mapped to a point on the earth's surface. To prepare the selected student dataset for storage within a geospatial database, addresses were first geocoded. Various geocoding methods were tested, most of which were expensive and time-consuming. Some of the geocoding tools available include 'geocod.io', geocoder wrappers from NodeJs, Ruby, and Python, and Tiger Geocoder extension for PostgreSQL. Geocoder wrapper tools implement geocoding through multiple sources such as Google, HERE, Yahoo, and MapBox but provide functions that abstract the programmer from having to use all the individual software. Most of these wrappers and geocoding tools were either unable to handle batch geocoding (geocoding large numbers of addresses) in a timely and cost-effective manner or had a limited number of addresses allowed for geocoding. Thus, the Tiger Geocoder extension was used for most large datasets while *geocod.io* was used for smaller datasets. Before geocoding

addresses, the Address Standardizer was used to ensure addresses were formatted and stored correctly.

```
SELECT    num, street, city, state, zip, zipplus
FROM      parse_address('25 Wizard of Oz, Walaford, KS 99912323') as a
;
```

(a) `parse_address` function

	num	street	city	state	zip	zipplus
1	25	Wizard of Oz	Walaford	KS	99912	323

(b) Normalized Result

```
SELECT    (a).num, (a).street, (a).city, (a).state, (a).zip, (a).zipplus
FROM      (
          SELECT    parse_address(addr.address) as a
          FROM      (
                    SELECT    concat(stu.street_line1, ', ', stu.street_line2, ' ',
                                     stu.street_line3, ', ',stu.city, ', ',stu.state,
                                     ', ',stu.postal_code) as address
                    FROM      student_addresses stu
                    ) as addr
          ) as p
;
```

(c) Parsing Addresses in the Test Database

Figure 3.1: Address Standardizer Usage

3.2.1 Address Standardizer

The Address Standardizer extension is a PostgreSQL extension that takes a single line address and parses it, with the help of three tables; the *rules* table, the *lex* table and the *gaz* table. The *rules* table provides the basic mapping rules, the *lex* table deals with alphanumeric input, and the *gaz* table is used to standardize place names. These tables are generated in the database during the installation of the extension. More details can be found in the PostGIS manual [7].

In the student database, addresses are provided within the columns: `street_line1` `street_line2`, `street_line3`, `city`, `state`, `postal_code`, and `country`. However,

many addresses in the student database are not formatted properly within these fields. The address_standardizer extension is used to normalize the addresses by parsing single addresses formed by concatenating the fields. Figures 3.1a and 3.1b show an example of normalizing an address with the extension and the result from the query. Figure 3.1 shows how addresses are normalized within the test database.

3.2.2 Tiger Geocoder

Tiger Geocoder is an extension for geocoding addresses into latitude and longitude coordinates. It comes with its own address normalizer and includes functions for reverse geocoding, geospatial indexing, as well as other geospatial functions which load and operate on geospatial census data. It is written to work with the Topologically Integrated Geographic Encoding and Referencing system (TIGER) released by the U.S Census Bureau, and is designed specifically for U.S addresses [7]. The address normalizing function within Tiger Geocoder depends on the Address Standardizer extension introduced in the previous subsection and works similarly.

Tiger Geocoder depends on data from the `tiger_data` schema which needs to be downloaded and set up. Shapefiles and lookup data can be downloaded for all states and territories in the U.S. Further information about the installation of the `tiger_data` are available in the PostGIS manual [7]. Disadvantages of using Tiger Geocoder for geocoding include the tasking setup process, slow batch geocoding time, and a significant amount of space required for storing TIGER data. The major advantage is that it is free, unlike other geocoders. For fewer addresses, alternative geocoders such as those from Google and HERE are sufficient. Once the TIGER data is set up, the table with the list of addresses can be geocoded. To geocode the `student_addresses` table, the columns `lat`, `lon`, `rating` and `new_address` were added to the table as shown in Figure 3.2. These columns were necessary for the

latitude, longitude, match rating, and normalized address fields generated by the geocoder.

```
ALTER TABLE       student_addresses
  ADD addid       SERIAL PRIMARY KEY,
  ADD lon         NUMERIC,
  ADD lat         NUMERIC,
  ADD new_address text,
  ADD rating      INTEGER
;
```

Figure 3.2: Adding Columns for Geocoder

After all addresses were geocoded, the `geom` column was added, and designated a *Point* geospatial datatype for the coordinate values. The `geom` column is essential for geospatial operations and can be designated any of the geospatial datatypes available in PostGIS. Tiger Geocoder's `geocode` function takes in the addresses as input and produces normalized addresses. The script for updating student addresses with the normalizer is shown in Figure 3.3.

```
UPDATE        student_addresses
SET           (rating, new_address, lon, lat)
              = ( COALESCE((g.geo).rating,-1), pprint_addy((g.geo).addy),
              ST_X((g.geo).geomout)::numeric(8,5), ST_Y((g.geo).geomout)::numeric(8,5) )
FROM
  (SELECT addid
      FROM student_addresses
      WHERE rating IS NULL ORDER BY addid) As a
      LEFT JOIN
        (SELECT addid, (geocode(address,1)) As geo
            FROM student_addresses As ag
            WHERE ag.rating IS NULL ORDER BY addid) As g ON a.addid = g.addid
WHERE a.addid = student_addresses.addid;
```

Figure 3.3: Geocoding Addresses with Tiger Geocoder

Pre-Institutional	Enrollment-Based	Post-Graduation
High School	CGPA	Career
HSGPA	Retention	Relocation
SAT / ACT scores	Time to degree	
	Course load	
	Housing (Off or On Campus)	

Table 3.1: Attributes of Students

3.3 Geospatial Queries and Joins

Geospatial databases are mostly regular databases extended to provide extra functionality. Hence, geospatial queries are very similar to regular queries for most databases but with added functionality to account for geospatial data structures and spatial relations. There are several geospatial databases or databases with geospatial extensions. As mentioned in Chapter 1, three main questions illustrate the types of information that can be extracted from the geospatial student database. These questions are dealt with in the subsequent subsections.

3.3.1 Do the attributes of students from a particular region provide information about the region?

In this context, student attributes refer to all attributes that could be associated with a student in the dataset. Attributes may be pre-institutional, post-graduation, or enrollment-based. Pre-institutional attributes are usually predetermined and consist of student characteristics that exist before matriculation. Examples of student attributes are provided in Table 3.1.

By using student attributes to provide information about geographic regions, the queries are a means of creating profiles on geographic regions based on these at-

```
SELECT      ceng.namelsad10 as census_name, ceng.countyfp10 as county_code,
            stu.gender, stu.ipeds_values_desc as ethnicity,
            stu.major_desc as major, stu.gpa,
            stu.nsemenrl as number_of_semesters_enrolled
FROM        student_data stu
LEFT JOIN   student_addresses addr
  ON        addr.unm_banner_id = stu.unm_banner_id
LEFT JOIN   census_shapefile ceng
  ON        st_contains(ceng.geom, addr.geom)
ORDER BY    ceng.name10
;
```

Figure 3.4: Attributes of Students within Census Tracts

tributes. In Structured Query Language (SQL), such queries are implemented by joining the tables consisting of geospatial information to the tables comprised of student information, and grouping them by geographic locations. Joining a geospatial table containing geographic locations to a table with student information requires a spatial join. More specifically, the join requires finding student addresses contained within particular regions. The `ST_Contains` function provided by PostGIS can handle such joins. As shown in Figure 3.4, simply selecting all relevant student attributes and census tracts from the spatially joined `student_data` and `census_shapefile` tables should provide relevant student data for each census tract. The query selects only few student attributes for brevity. Counties, school districts, and other geographic regions could substitute census tracts.

Another way of relating student attributes to geographic regions is through aggregate results such as averages, percentages, and counts within each geographic region. The example in Figure 3.5 shows female and male percentages, average GPAs, and time-to-degree for students within each county from the dataset.

```
SELECT      ceng.namelsad10 as county,
            round((count(DISTINCT case WHEN  stu.gender = 'F' THEN
              stu.unm_banner_id END) * 100)::NUMERIC /
              count(DISTINCT stu.unm_banner_id), 2) as female_percentage,
            round((count(DISTINCT case WHEN  stu.gender = 'M' THEN
              stu.unm_banner_id END) * 100)::NUMERIC /
              count(DISTINCT stu.unm_banner_id), 2) as male_percentage,
            round(avg(stu.gpa), 2) as average_gpa,
            round(avg(stu.nsemenrl), 2) as average_number_of_semesters_enrolled
FROM        student_data stu
LEFT JOIN   student_addresses addr
  ON        addr.unm_banner_id = stu.unm_banner_id
LEFT JOIN   county_shapefile ceng
  ON        st_contains(ceng.geom, addr.geom)
WHERE       ceng.namelsad10 IS NOT NULL
GROUP BY    ceng.namelsad10
ORDER BY    count(DISTINCT stu.unm_banner_id) desc
;
```

(a) Aggregate query of student attributes per county

	county	female_percentage	male_percentage	average_gpa	average_number_of_semesters_enrolled
1	Bernalillo County	61.76	38.24	3.43	8.8
2	Santa Fe County	68.09	31.91	3.57	8
3	Sandoval County	60.47	39.53	3.54	8.68
4	Valencia County	63.16	36.84	3.53	7.84
5	Doña Ana County	61.54	38.46	3.6	7.5
6	Taos County	100	0	3.66	7.5
7	San Juan County	50	50	3.59	8
8	Rio Arriba County	33.33	66.67	3.45	8.33
9	Los Alamos County	66.67	33.33	3.78	7.67
10	McKinley County	66.67	33.33	3.35	7.5
11	San Miguel County	25	75	3.47	9.5

(b) Results of Aggregate Query

Figure 3.5: Aggregate Query for Students within Counties

3.3.2 Do socioeconomic data based on students' geographic origins relate to student outcomes and choices?

This question follows the same concept introduced in the previous section. Students are spatially joined to the census via the `ST_Contains` function. However, queries require a different point of view, where data from geographic regions are analyzed to figure out how regional characteristics affect students who live within them. Thus, both socioeconomic data from census data, and student attributes from

DP02	DP03	DP04	DP05
Average Family Size	Income $200k or more	paying 35% or more GRAPI	Population Under 5 years
Average Household Size	With Health Insurance Coverage		Hispanic or Latino (of any race)
High School Attainment	No Health Insurance Coverage		American Indian
Native Population (born within the U.S)	Course load		
Foreign Population			

Table 3.2: Census Socioeconomic Data On Census Tracts

the Student Data Mart are selected in the queries. Consider as an example, a query on the relationship between average income within geographic regions and student GPAs. Average income is a socioeconomic factor, and GPA is the student attribute in this case. The socioeconomic data provided by the census is categorized under five major data profiles as listed below.

- DP01: Profile of General Population and Housing Characteristics
- DP02: Selected Social Characteristics in the United States
- DP03: Selected Economic Characteristics
- DP04: Selected Housing Characteristics
- DP05: ACS Demographic and Housing Estimates

For the test database, various socioeconomic characteristics were selected from four of the data profiles, as shown in Table 3.2. To figure out how socioeconomic data

on census tracts relates to student outcomes, queries involved selecting these characteristics from the `census_socioecons` table, as well as, student attributes used to measure student outcomes. The `census_socioecons` table was spatially joined to the students via the `census_shapefile` and `student_addresses` tables. The Entity Relationship Diagram (ERD) in Chapter 2 Figure 2.1, shows the relationships between the tables. Figure 3.6 shows an example query used to derive some characteristics from the `census_socioecons` table and Figure 3.7 shows a query with aggregate functions of averages for counties.

```
SELECT      ceng.namelsad10 as census_name, stu.unm_banner_id as student_id,
--  student attributes such as gpa and # of semesters enrolled till graduation
            stu.gpa, stu.nsemenrl as number_of_semesters_enrolled,
--  socioeconomic data from census data
            socio.avg_family_size, socio.below_fpl_all as percentage_below_fpl,
            socio.hs_attain_25 as high_school_attainment_at_age_25,
            socio.hisp_latino_any as percentage_of_hispanics
FROM        student_data stu
-- join students to their addresses
LEFT JOIN   student_addresses addr
  ON        addr.unm_banner_id = stu.unm_banner_id
-- spatially join student addresses to census tracts
LEFT JOIN   census_shapefile ceng
  ON        st_contains(ceng.geom, addr.geom)
-- join census tracts to socioeconomic census data
LEFT JOIN   census_socio_econs socio
  ON        socio.geoid = ceng.geoid10
-- associate census tract data for prior 5 yrs to students' matriculation yrs
WHERE       ((socio.year >= 2011
  AND       socio.year <= 2015
  AND       socio.year = stu.year_admitted)
  OR        (socio.year > 2015
  AND       stu.year_admitted = 2015)
  OR        (socio.year < 2011
  AND       stu.year_admitted = 2011))
  AND       socio.geotype = 'Census Tract'
ORDER BY    ceng.name10
;
```

Figure 3.6: Socioeconomic Characteristics for Students within Census Tracts

```
SELECT      ceng.namelsad10 as county,
            round(avg(socio.native), 2) as native_percentage,
            round(avg(socio.with_health_insurance), 2) as percentage_with_health_insurance,
            round(avg(socio.hs_attain_25), 2) as high_school_by_age_25,
            round(avg(socio.hisp_latino_any), 2) as hispanic_percentage,
            round(avg(stu.gpa), 2) as average_gpa,
            round(avg(stu.nsemenrl), 2) as average_semesters_enrolled
FROM        student_data stu
  -- join students to their addresses
LEFT JOIN   student_addresses addr
  ON        addr.unm_banner_id = stu.unm_banner_id
-- spatially join student addresses to census tracts
LEFT JOIN   county_shapefile ceng
  ON        st_contains(ceng.geom, addr.geom)
-- join census tracts to socioeconomic census data
LEFT JOIN   census_socio_econs socio
  ON        socio.geoid = ceng.geoid10
-- associate census tract data for prior 5 yrs to students' matriculation yrs
WHERE       ((socio.year >= 2011
  AND       socio.year <= 2015
  AND       socio.year = stu.year_admitted)
  OR        (socio.year > 2015
  AND       stu.year_admitted = 2015)
  OR        (socio.year < 2011
  AND       stu.year_admitted = 2011))
  AND       socio.geotype = 'County'
GROUP BY    ceng.namelsad10
ORDER BY    ceng.namelsad10
;
```

(a) Aggregate Query of Socioeconomic Characteristics per County

	county	native_percentage	percentage_with_health_insurance	high_school_by_age_25	hispanic_percentage	average_gpa	average_semesters_enrolled
1	Bernalillo County	89.18	83.42	24.01	48.01	3.48	8.15
2	Chaves County	86.5	80.1	27.1	52	3.98	8
3	Cibola County	95.7	71.6	43	36.7	3.36	8
4	Colfax County	95.85	83.5	33.9	47.65	3.38	6
5	Doña Ana County	83.43	78.01	22.11	66.01	3.61	7.52
6	Eddy County	95.23	84.97	31.93	44.5	3.96	6.67
7	Grant County	95.3	85.7	27.8	48.5	4	8
8	Lea County	85.7	78.4	28.7	51.2	3.17	8
9	Lincoln County	93.8	82.5	26	30.1	3.83	8
10	Los Alamos County	90.3	94.9	11.28	15.2	3.97	6.8
11	Luna County	84.1	77	31.9	61.7	3.78	8
12	McKinley County	97.81	63.46	31.86	13.65	3.35	8

(b) Results of Aggregate Query

Figure 3.7: Aggregate Query of Socioeconomic Characteristics Within Geographic Regions

In determining how socioeconomic factors affect students, aggregate functions play the most relevant role. A variety of aggregate functions provide information when classification systems are introduced. A classification system, in this case, refers to a means of classifying students by socioeconomic categories. An example would be a binary classification system for determining whether a geographic region is considered high income. Consider for the data profile, DP03, percentages of people with an income of $200,000 or more. Using the average as a pivoting point, com-

munities where percentages are equal or greater than the average are assumed to be high income and assigned a value 1, while 0 is assigned to low-income communities. The code and results for the classification query are provided in Figure 3.8. Student outcomes such as average GPAs or time-to-degree can be measured against these two classifications. Several classification methods can be introduced into the database and used with regression models or machine learning algorithms.

```
SELECT      socio.year, round(yr.avg_income, 2),
            round((sum(CASE WHEN socio.income_over_200k >= yr.avg_income
              THEN 1 END)::NUMERIC * 100) / count(socio.income_over_200k), 2)
              as wealthy_census_tracts,
            round((sum(CASE WHEN socio.income_over_200k < yr.avg_income
              THEN 1 END)::NUMERIC * 100) / count(socio.income_over_200k), 2)
              as less_wealthy_census_tracts,
            round(avg(stu.gpa), 2) as avg_gpa, round(avg(stu.nsemenrl), 2)
              as avg_enrollement_terms
FROM        student_data stu
-- join students to their addresses
LEFT JOIN   student_addresses addr
  ON        addr.unm_banner_id = stu.unm_banner_id
-- spatially join student addresses to census tracts
LEFT JOIN   county_shapefile ceng
  ON        st_contains(ceng.geom, addr.geom)
-- join census tracts to socioeconomic census data
LEFT JOIN   census_socio_econs socio
  ON        socio.geoid = ceng.geoid10
-- associate census tract data for prior 5 yrs to students' matriculation yrs
LEFT JOIN   (SELECT year, avg(income_over_200k) as avg_income
             from census_socio_econs
             GROUP BY year) yr
  ON         yr.year = socio.year
WHERE       ((socio.year >= 2011
  AND       socio.year <= 2015
  AND       socio.year = stu.year_admitted)
  OR        (socio.year > 2015
  AND       stu.year_admitted = 2015)
  OR        (socio.year < 2011
  AND       stu.year_admitted = 2011))
  AND       socio.geotype = 'County'
GROUP BY    socio.year, yr.avg_income
ORDER BY    socio.year
;
```

(a) Query for Student Outcomes in High Income Census Tracts

	year	round	wealthy_census_tracts	less_wealthy_census_tracts	avg_gpa	avg_enrollement_terms
1	2009	2.95	14.29	85.71	3.34	9.38
2	2010	3.07	14.46	85.54	3.34	9.39
3	2011	3.51	8.43	91.57	3.35	9.39
4	2012	2.63	90.48	9.52	3.53	7.96
5	2013	2.03	89.51	10.49	3.73	6

(b) Results of Classification Query

Figure 3.8: Aggregate Querying using Classification

3.3.3 What geospatial information can be gathered about students and academic institutions?

This line of query primarily involves spatial-centric information such as students' proximity to university campuses, availability of public transportation, proximity to disaster-prone areas, and availability of certain infrastructure. Geospatial information such as distance to the university campuses can easily be provided through geospatial querying of student and campus addresses. However, other geospatial queries may require additional layers of geospatial information.

This highlights a special trait of geospatial databases. Just as tables containing different aspects of student information can be added to a database, additional layers of geospatial information can be added to a database. For a map showing the various cities within New Mexico, a layer of traffic and highway information could be added to the map. Furthermore, additional layers of zip code boundaries, hospitals, coffee shops, hotels, public buildings, population densities, and other information could be provided as labels, shapes, dots, and colors on the map. This drastically expands the possible approaches by which student data can be analyzed. However, care must be taken to limit geospatial queries to factors that could actually affect a student or inform an institution about a lack of particular relevant services and infrastructure.

Shapefiles of federal highways, roads, public facilities, landmarks, urban areas and school districts were imported into the test database to demonstrate other potential queries with a variety of geospatial data. The ERD is displayed in Figure 3.9. With all these shapefiles, so many questions can be asked about students and UNM, including the following:

- *Which students live closest to the Westgate Library?*
- *Are there any landmarks close to UNM?*

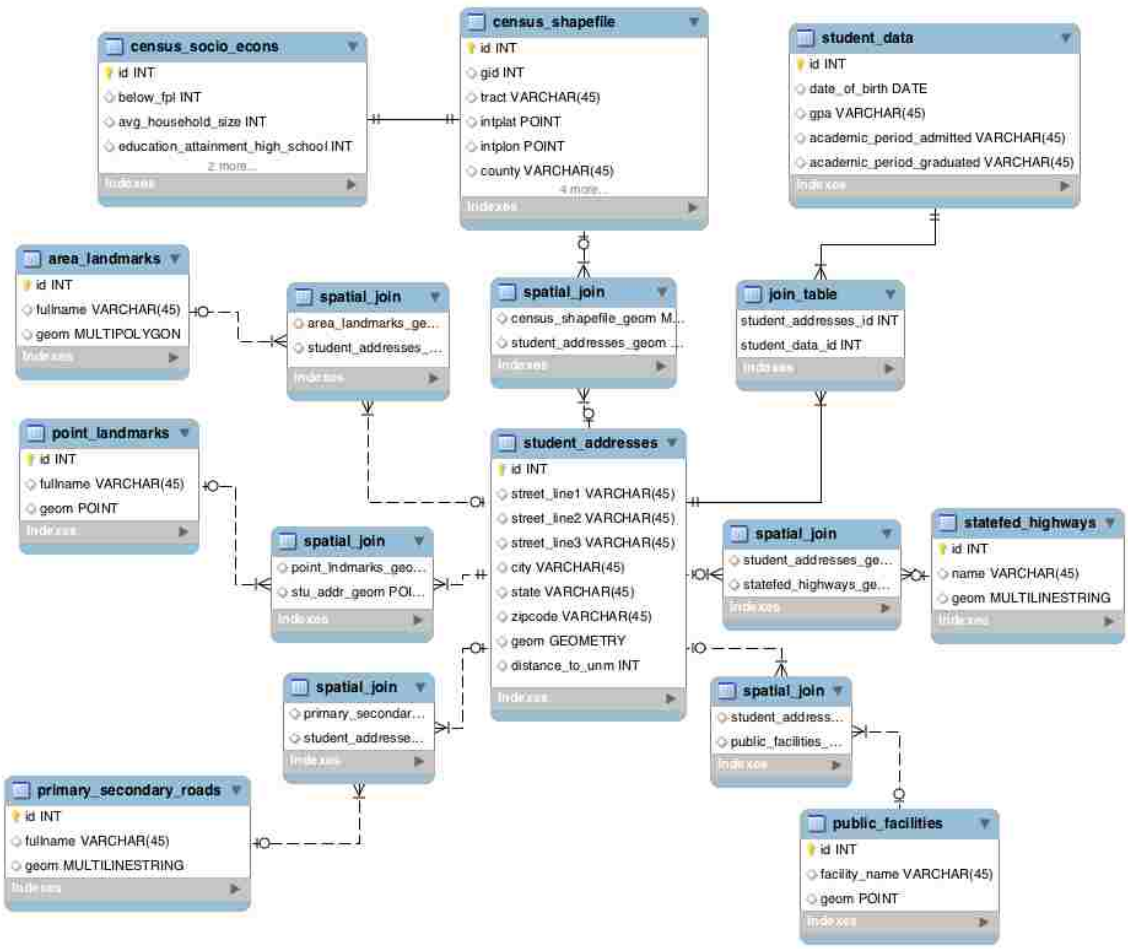


Figure 3.9: An Entity Relationship Diagram Showing Possible Geospatial Additions to A Student Database

- *What are some socioeconomic characteristics of UNM's census tract?*
- *Which public facilities are available to students living downtown*
- *Are there good roads connecting students living in the downtown dormitory to UNM for school buses?*

To demonstrate some of the geospatial functions available in PostGIS, the first question is illustrated in Figure 3.10. Two geospatial functions are used in the query,

`ST_Distance` and `ST_DWithin`. `ST_Distance` is used to measure the distance between two geospatial objects. In this case, it is used to measure the distance between student addresses and the Westgate Library. While the unit provided in the results is in meters. PostGIS provides other functions such as the `ST_Transforms` function for converting the measurements into other desired units. The `ST_Dwithin` in the query is used to join the `public_facilities` table to the `student_addresses` table based on all students within the given distance. Several other geospatial functions are provided in the PostGIS manual for geospatial operations [7].

```
SELECT      DISTINCT round((random() * 1000000)::NUMERIC) as student_id,
            -- random numbers are generated for student IDs
            pf.facilityna, -- name of facility
            pf.facilityty, -- type of facility
            round(st_distance(stu.geom, pf.geom)::NUMERIC, 2) as distance_srid_units
            -- srid_units are the units of measurement resulting from st_distance
            -- they can be converted to miles, meters or other units based on the geom srid type
FROM        public_facilities pf
LEFT JOIN   student_addresses stu
  ON        st_dwithin(stu.geom, pf.geom, 2100000)
WHERE       pf.facilityna = 'WESTGATE'
  AND       pf.facilityty = 'LIBRARY'
GROUP BY    pf.facilityna, pf.facilityty, st_distance(stu.geom, pf.geom)
ORDER BY    distance_srid_units
LIMIT       10
;
```

(a) Query of Students Living Closest to Westgate Library

	student_id	facilityna	facilityty	distance_srid_units
1	711003	WESTGATE	LIBRARY	2096412.31
2	830962	WESTGATE	LIBRARY	2096413.93
3	980081	WESTGATE	LIBRARY	2096413.94
4	477483	WESTGATE	LIBRARY	2096414.19
5	675955	WESTGATE	LIBRARY	2096414.19
6	39112	WESTGATE	LIBRARY	2096414.51
7	261276	WESTGATE	LIBRARY	2096414.51
8	749797	WESTGATE	LIBRARY	2096414.51
9	210591	WESTGATE	LIBRARY	2096414.52
10	232735	WESTGATE	LIBRARY	2096414.52

(b) Result of Distance Query

Figure 3.10: Students Living Close to the Westgate Library

Chapter 4

Analyzing the Data

Extensive analyses show various relationships between student backgrounds and performance when socioeconomic data is related to student data. While more extensive machine learning and data analysis measures could be applied to the data, the dataset for the study was limited and merely serves to introduce basic data analysis on student data. In this section, two methods used to analyze the data are demonstrated: Data visualization with Tableau and computations using Python.

The Census Bureau has been very concerned with the educational attainment of students, especially from Hispanic, American Indian, and African American communities [12]. This is mainly due to the fact that these ethnic groups have typically experienced lower retention and graduation rates [12]. As New Mexico has a large concentration of Hispanic and American Indian populations in various geographic regions, most of the visualizations focused on students from these geographic regions. This is also due to the lack of sufficient numbers of students from communities with other ethnic backgrounds needed to provide relevant statistical data.

4.1 Data Visualization

At UNM's Office of Institutional Analytics, Tableau is used extensively to visualize data into various graphs, charts, and maps that make it easier to recognize patterns and analyze the data. With Tableau, dashboards are seamlessly created by connecting to a data source and selecting the desired parameters or database columns for illustration purposes. Tableau also allows for most basic queries and joins, including geospatial features, making it a suitable tool for geospatial data visualization.

Using various charts, maps and graphs with colors, the relationship between student attributes and socioeconomic data are more pronounced, and further analysis can be conducted. Some examples of simple data visualizations that show basic patterns and characteristics of students in different geographic areas within New Mexico are explored in this section. A dataset of undergraduate students at UNM who live in New Mexico and were enrolled during fall 2016 was selected. For demonstration purposes, the illustrations are based on the same three questions presented in Chapters 1 and 3.

Do the attributes of students from a particular region provide information about the region?

Suppose a university needed to know whether students from various census tracts experience large disparities in student outcomes. An optimal data visualization would show the time-to-degree, GPAs, and HSGPAs of students from these census tracts. As illustrated in Figure 4.1, within the state of New Mexico, census tracts with the highest average HSGPAs tend to maintain this momentum through college. However, time-to-degree seems to be independent of GPAs within census tracts. Besides student outcomes, other attributes of students such as ethnicity, age distribution, choice of housing, and student majors could be observed within census tracts.

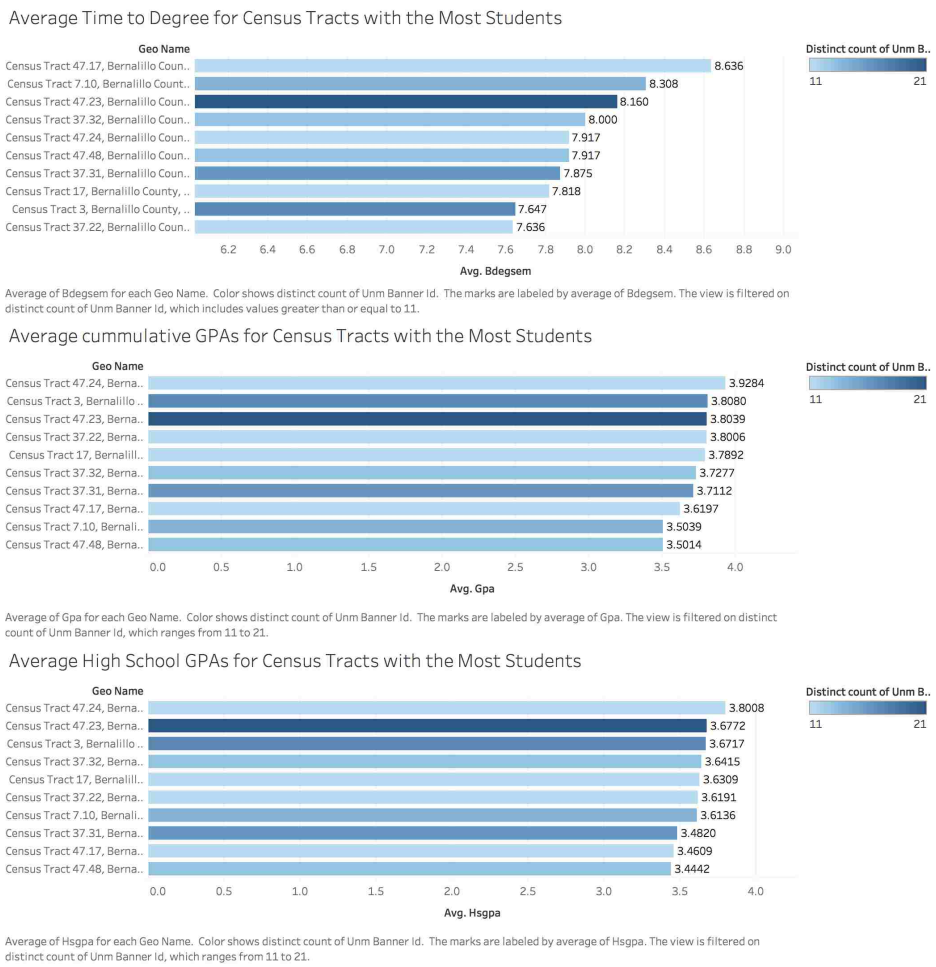
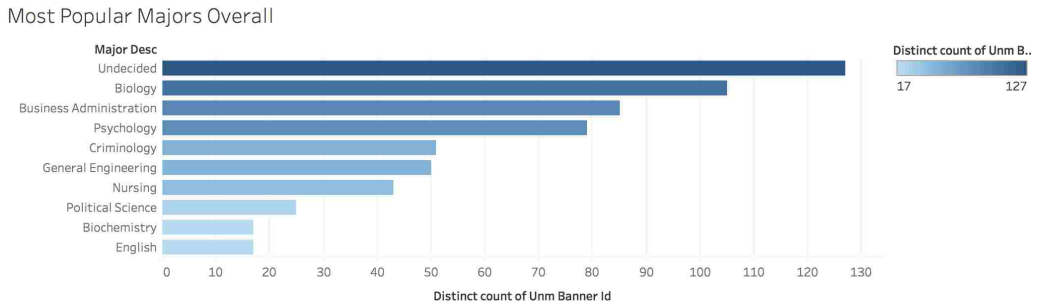


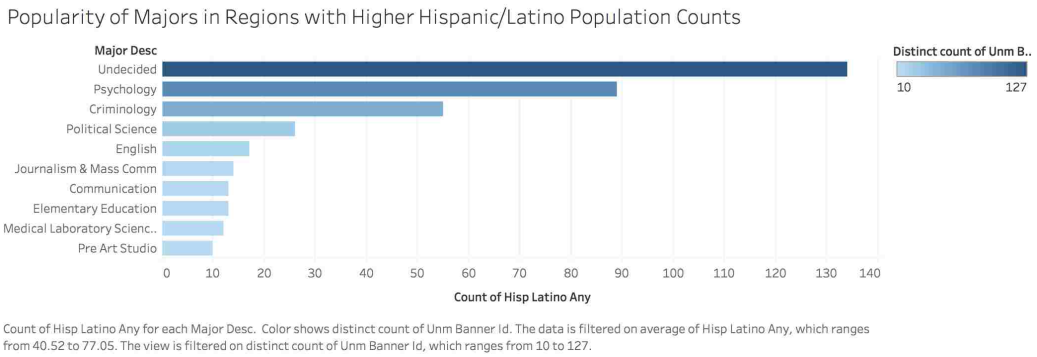
Figure 4.1: Student Attributes for Census Tracts with the Highest Populations of UNM Students

Do socioeconomic data based on students' geographic origins relate to student outcomes and choices?

One interesting observation was how geographic regions that students lived in, affected student GPAs, time-to-degree and choice of majors and housing. The most popular majors at UNM within the selected dataset in order of magnitude were Biology, Business Administration, Psychology and Criminology. Psychology and Business Administration switch places once their distribution across census tracts is taken



(a) Popularity of Majors at UNM for students who live in New Mexico



(b) Popularity of Majors in Regions with More Hispanics and Latinos

Figure 4.2: The popularity of Majors within the University

into consideration. However, the popularity of Biology and Business Administration drastically drops when taking into account students from geographic regions with higher concentration of Hispanics. Students from geographic regions with higher percentages of Hispanics chose Psychology, Criminology, and Political Science while students from regions with fewer Hispanics chose Biology and Business Administration. In geographic regions with more natives (people born in the United States), most students had undecided majors. Biology, Psychology and Business Administration were also the most popular majors among such students, while Criminology and Elementary Education were popular among students with higher concentrations of non-natives. The pivoting point to determine whether or not a geographic region was

deemed to have a higher or lower concentration of a given attribute was the average. For example, concerning students from geographic regions with higher concentration of Hispanics, a filter was set for the percentage of Hispanics within the specific regions to be greater than the average. For brevity, Figure 4.2 shows the distribution of majors among all students in the dataset and students from geographic regions with higher Hispanic/Latin concentrations, while the rest of the figures are provided in Appendix C.

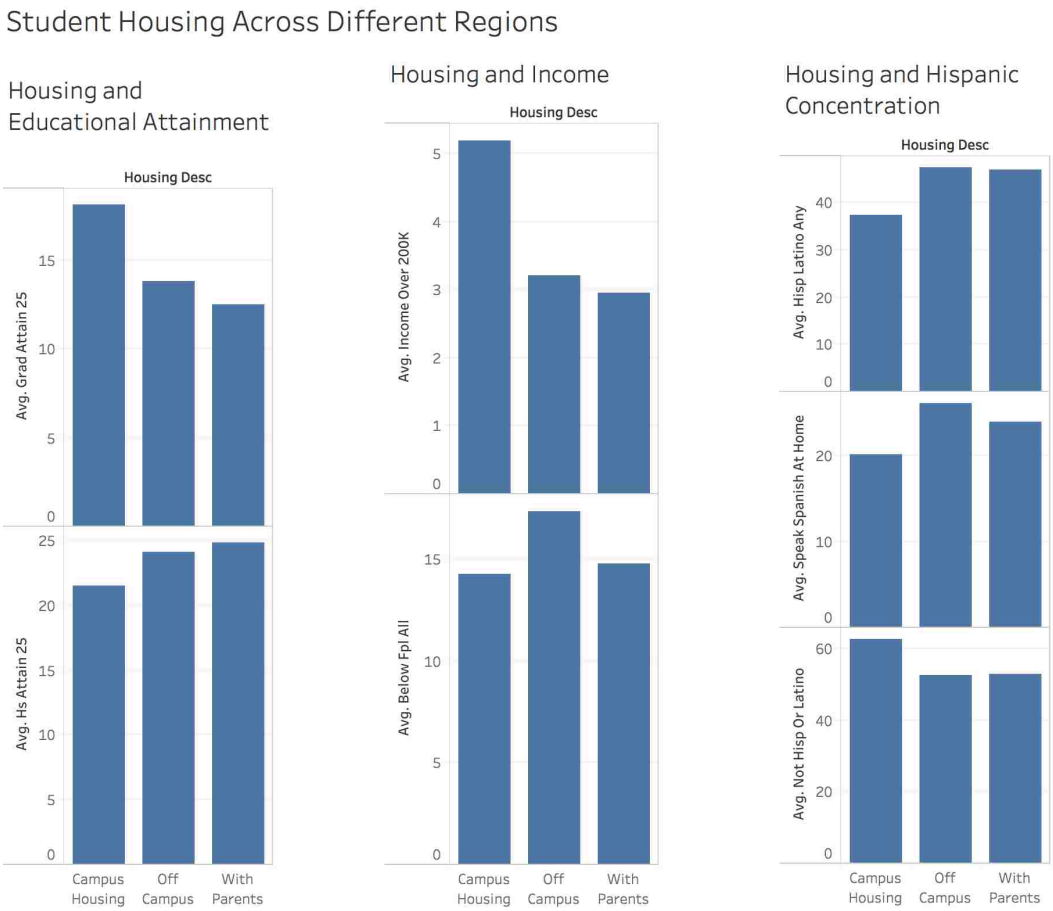


Figure 4.3: Student Housing

Housing characteristics were also different among students from different backgrounds. Students from regions with higher percentages of Hispanics and people

below the Federal Poverty Level (FPL) preferred to live off-campus or with their parents. The inverse was the case for students from populations with higher educational attainment percentages and higher percentages of wealthy residents as shown in Figure 4.3. GPA and time-to-degree also showed slight differences when plotted against students from different backgrounds. Students from regions with higher percentages of Hispanics, higher percentages of people below the FPL and lower percentages of health insurance coverage, took a longer time to graduate and had relatively lower GPAs overall. However, the differences were not large enough to make these attributes predictive indicators for all students

What geospatial information can be gathered about students and academic institutions?

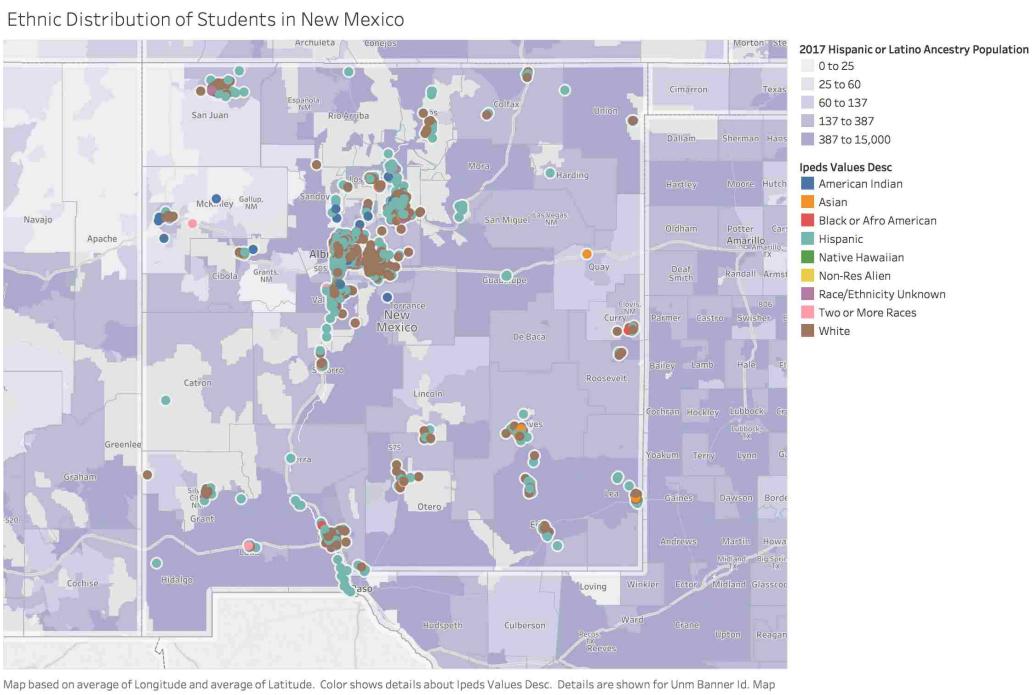


Figure 4.4: Ethnic Distribution of Students in New Mexico

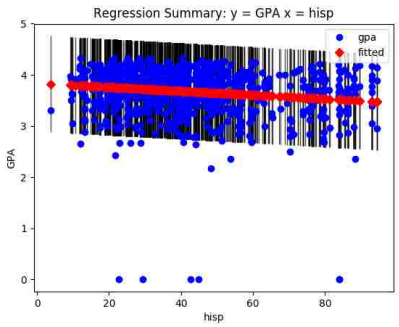
With Tableau, the students' ethnic distribution across New Mexico could be illus-

trated as shown in Figure 4.4. The legend binds colors to ethnicity and population density. From the map, it is apparent that most students are White, Hispanic, and American Indian with the highest concentration of students within and around Bernalillo County. This makes sense since the university is located within the area. By hovering the pointer over a student on the map, Tableau provides detailed information about the student based on selected parameters. Furthermore, various census datasets are available in Tableau as data layers. In Figure 4.4 for example, a data layer showing the 2017 Hispanic or Latino ancestry populations is depicted by different shades of purple for various census blocks within the state of New Mexico.

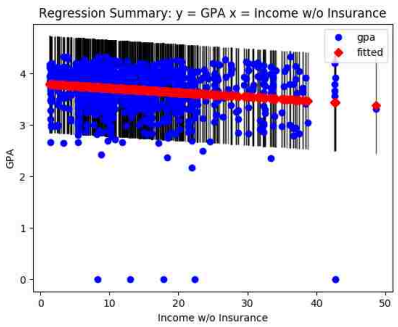
4.2 Statistical Analysis with Python

A possible application of a geospatial student database is predictive analysis. So far, geospatial querying has been introduced as a means to extract socioeconomic and geospatial information about students and their environments. Taking it a step further, the information could be analyzed with regression models and adapted into predicting student outcomes and choices. To test the possibility of regression modeling and predictive analysis using geospatial student data, census socioeconomic features of students' communities were plotted against selected measures of student outcomes. The two main selected measures were the time-to-degree and GPA of students. A simple regression model known as the Ordinary Least Squares (OLS) regression model was used. The code is provided in Figure 4.6. Some socioeconomic factors were found to affect student outcomes. For example, students from Hispanic communities took longer to graduate on the average and had lower GPAs as shown in Figure 4.5a. However, as seen in the graphs, the regression lines were not a good predictive measure overall because the differences between students from communities with higher concentrations of Hispanics and those with lower concentrations were

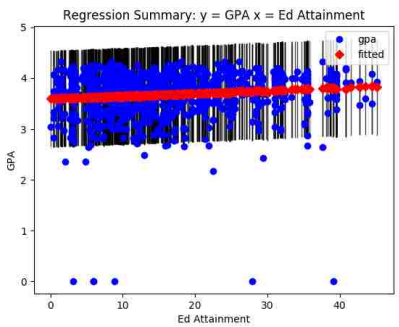
small. Since there is a relation, however, machine learning algorithms and specialized regression models could produce more robust predictions of incoming students' outcomes based on historical socioeconomic values of student backgrounds.



(a) Hispanic Concentration in Communities Plotted Against GPA



(b) Low Health Insurance Coverage Plotted Against GPA



(c) Communities with Higher Educational Attainment Percentages Plotted Against GPA

Figure 4.5: Socioeconomic Features and Their Influence on Student GPAs

For regression analysis with Python, a number of packages were imported, including Pandas, GeoPandas, Statsmodels and SQLAlchemy. Pandas and GeoPandas are open source Python libraries that provide data structures and data analysis tools. These packages are comparable to working with R but provide less functionality with data analysis. R is a programing language designed specifically for statical computing, and also provides packages for statistical analysis on geospatial

data. Two powerful data structures made available by the Pandas library are Series and DataFrames. A Series is a one dimensional data structure of arrays while a DataFrame is a two dimensional data structure with columns of potentially different data types. These two data structures come with various methods and attributes that make data analysis with Pandas relatively simple. GeoPandas is an extension of the Pandas package with geospatial functionality and provides the *GeoSeries* and *GeoDataFrame* data types as extensions to Series and DataFrame types. These data types allow for easy computational operations and analysis with Python as shown in the code in Figure 4.6. As shown in the code, SQLAlchemy is used to connect to the test database. Pandas provides the `read_sql_table` method for importing the dataset into a Python DataFrame. Using the `Statsmodels.api` package, statiscal analysis such as the Ordinary Least Squares algorithm can be run on the data.

subset_regression.py

```python
import pandas as pd
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
from sqlalchemy import create_engine

# connect to test database
engine = create_engine('postgresql://edwinagbenyega@localhost:5432/thesis')

# pull data from subset table (a table containing student data joined to census data for
# data analysis
df = pd.read_sql_table(table_name="subset", con=engine)

# regression function for displaying results from and OLS regression model
def reg_m(y, x):
    ones = np.ones(len(x[0]))
    X = sm.add_constant(np.column_stack((x[0], ones)))
    for ele in x[1:]:
        X = sm.add_constant(np.column_stack((ele, X)))
    results = sm.OLS(y, X).fit()
    return results


# plotting the results of the regression model onto a graph
def plt_graph(xlabel, ylabel, title, results, fname):
    fig, ax = plt.subplots()
    fig = sm.graphics.plot_fit(results, 0, ax=ax)
    ax.set_ylabel(ylabel)
    ax.set_xlabel(xlabel)
    ax.set_title(title)
    plt.savefig('./graphs/' + fname + '.png')

# variables containing socioeconomic factors and student attributes for regression analyses
df2 = df[df['bdegsem'] >= 6]
y1 = df2['bdegsem']
y2 = df2['gpa']
x1 = [df2['hisp_latino_any']]
x2 = [df2['hisp_latino_any'], df2['native'], df2['foreign_born']]
x3 = [df2['no_health_insurance']]
x4 = [df2['college_grad'], df2['hs_attain_25'], df2['grad_attain_25']]

# dictionaries for possible variable combinations of socioeconomic factors and student attributes
y = {'BDEGSEM': y1, 'GPA': y2}
x = {'hisp': x1, 'Racial Distribution': x2, 'Income w/o Insurance': x3, 'Ed Attainment': x4}

# df2['income_over_200k'], df2['below_fpl_all'],

# running regression model for all the factors included in the x and y dictionaries
count = 0
for i,j in y.items():
    for k,l in x.items():
        count += 1
        print()
        title = "Regression Summary: y = " + i + " x = " + k
        reg = reg_m(j, l)
        print(title)
        print('*'*100)
        print(reg.summary())
        plt_graph(k, i, title, reg, i + str(count))
```


Figure 4.6: Python Code for OLS Regression Modeling

Chapter 5

Future Work / Conclusion

5.1 Geospatial Profiling as a Potential Substitute for Census Data

According to a study of decennial census undercounts by the PricewaterhouseCoopers consulting firm, the Census undercounted the population of New Mexico by an estimated 35,988 people, or 9.4% of the population in 2000 [13]. This trend continued in subsequent years and the undercounted estimates ultimately cost the state over $100 million in Federal funding between 2002 and 2012 [14]. Data compiled by the Bureau of Business and Economic Research(BBER) at UNM suggests that the effect of the undercount may have been even more severe [14]. It comes to question then whether the Census Bureau's data is an appropriate source for examining student backgrounds.

The Census Bureau is a rich source of data, statistical reports and analyses on various characteristics of people within the U.S. With regards to school enrollment and other educational attainment characteristics, the Census Bureau depends on a vari-

ety of surveys and data collection programs. These mainly consists of the American Community Survey (ACS), Current Population Survey (CPS), Survey of Income and Program Participation (SIPP), Small Area Income and Poverty Estimates (SAIPE) Program, and other smaller survey bodies [15, 12].

While these programs make an effort to provide information based on answers derived from surveys, factors such as sampling and non-sampling errors may account for certain levels of inaccuracy [15]. Factors such as respondents' interpretation of questions, cooperation, and accuracy of answers are typical causes of non-sampling errors [15]. Quality control measures and adjusted weighted procedures of the final estimates are carried out to reduce these errors to acceptable error margins [15]. However, an alternative to depending on census data could be historical student data.

Arguably, statistical data made available by associating students with their geographic origins could provide information on geographic regions based on student characteristics. For example, by associating the average GPAs, retention rates, gender, ethnicities and other attributes of students with their addresses, more accurate and substantial data could be associated with various geographic regions. This can be achieved by exploiting geospatial relationships of historical data regarding students and where they lived. Also, geospatial querying could provide more valuable information such as the proximity of students to various facilities, percentages of students within walking distance to their educational institutions, concentration of students in particular regions per race, gender, and major.

5.2 Framework for Generating Statistical and Analytical Reports

Incorporating geospatial functionality into student databases opens up a myriad of possibilities. By adding various layers of geospatial data, different aspects of students and the geographic regions they live in are made available for analysis. As shown in Chapter 3, shapefiles with information about federal highways, urban and rural areas, landmarks, and public facilities provide information about the communities students live in and the resources available. Additional layers of spatial data from sources such as Google Maps and *data.gov* could also be integrated for further analysis. Since different parts of the U.S have different geographic characteristics, an interesting expansion of the research is to compare statistical geospatial data from various universities in different geographic regions. In order to achieve this, a future direction would involve creating a portable framework that generates statistical information based on Census Bureau socioeconomic data and student data.

The framework would have predefined variables that could be mapped onto fields from any educational institution's database. Once educational institutions input these fields, the framework would compare historical data with those of other institutions of similar backgrounds and produce predictive geospatial reports and analytics. The framework would include the ability to generate a PostgreSQL/PostGIS geospatial database based on student addresses. It would also use geospatial and machine learning functions from Python packages such as *Pandas/Geopandas* and *Sci-kit Learn.*

5.3 Conclusion

The U.S Census Bureau gathers data from all over the U.S for policymaking and allocation of funds to various communities [16]. Using geospatial databases as a means to associate socioeconomic census data with student data is a relatively new concept. Thus, there are a lot of unexplored analytical and predictive models. The thesis explored three broad areas of incorporating geospatial data into data analytics for HEIs. The first area considered how attributes of students living in a geographic region could be used to characterize the geographic region. The second explored how socioeconomic data from the Census Bureau could describe the students who live in a given geographic region. The third area dealt with geospatial data from other external sources and how they related to students and educational institutions. These are broad topics that illustrate the benefits of incorporating geospatial data into a student database. However, the relationships introduced by a geospatial student database are by no means exhausted in this thesis. While the study mentioned some problems worth addressing, there are still several unexplored questions.

Geospatial databases enable queries on 3D and 4D data. This means that geospatial analytics are not restricted to geographic regions but also encompass 3D and 4D spatial concepts. Therefore, researchers could explore topological relationships between university campuses and students. An example is the effect of landscape, architecture, and spatial dynamics of where students study (such as campuses, classrooms and libraries) on student outcomes. Geospatial databases also enable specialized analytics on students and classes. These include how student outcomes are influenced by the distribution of students within a classroom of a given size or the effect of class schedules on students with varied proximity to the school and accessibility to transportation.

Potential advances include expanding the dataset to consist of several universities

and creating a portable framework that other universities could apply to their data. By creating a framework for multiple universities, universities could share statistical geospatial data and learn more about incoming students from different geographic origins. Student advisors could help students align their courses and class schedules better, based on comparing statistical geospatial information. Projected statistical findings could also help identify discrepancies in student outcomes. Thus, geospatial student databases yield several benefits to HEIs and deserve further studies as an upcoming field of student analytics.

Appendices

Appendix A

Selected Fields from Census Data Mapped to Test Database Fields

```
h = {
    "GeoType": "geotype",
    "Id": "full_geoid",
    "GeoID": "geoid",
    "GeoName": "geo_name",
    "Year": "year",
    "Type": "type",
    "EDUCATIONAL ATTAINMENT - Percent bachelor's degree or higher": "bach_attain",
    "EDUCATIONAL ATTAINMENT - Percent high school graduate or higher": "hs_attain",
    "EDUCATIONAL ATTAINMENT - Population 25 years and over - Associate's degree": "assoc_attain_25",
    "EDUCATIONAL ATTAINMENT - Population 25 years and over - Bachelor's degree": "bach_attain_25",
    "EDUCATIONAL ATTAINMENT - Population 25 years and over - Graduate or professional degree": "grad_attain_25",
    "EDUCATIONAL ATTAINMENT - Population 25 years and over - High school graduate (includes equivalency)": "hs_attain_25",
    "EDUCATIONAL ATTAINMENT - Population 25 years and over - Less than 9th grade": "under_grade9_25",
    "EDUCATIONAL ATTAINMENT - Population 25 years and over - Some college, no degree": "no_deg_25",
    "GROSS RENT AS A PERCENTAGE OF HOUSEHOLD INCOME - Occupied units paying rent (excluding units where GRAPI cannot be computed) -
    35.0 percent or more": "rent_over_35p",
    "HEALTH INSURANCE COVERAGE - Civilian noninstitutionalized population - No health insurance coverage": "no_health_insurance",
    "HEALTH INSURANCE COVERAGE - Civilian noninstitutionalized population - With health insurance coverage": "with_health_insurance",
    "HISPANIC OR LATINO AND RACE - Total population - Hispanic or Latino (of any race)": "hisp_latino_any",
    "HISPANIC OR LATINO AND RACE - Total population - Not Hispanic or Latino": "not_hisp_or_latino",
    "HOUSEHOLDS BY TYPE - Average family size": "avg_family_size",
    "HOUSEHOLDS BY TYPE - Average household size": "avg_household_size",
    "INCOME AND BENEFITS - Total households - $200,000 or more": "income_over_200k",
    "LANGUAGE SPOKEN AT HOME - Population 5 years and over - Spanish": "speak_spanish_at_home",
    "PERCENTAGE BELOW FPL - 18 to 64 years": "below_fpl_18to64",
    "PERCENTAGE BELOW FPL - All families": "below_fpl_fam",
    "PERCENTAGE BELOW FPL - All people": "below_fpl_all",
    "PLACE OF BIRTH - Total population - Foreign born": "foreign_born",
    "PLACE OF BIRTH - Total population - Native": "native",
    "SCHOOL ENROLLMENT - Population 3 years and over enrolled in school - College or graduate school": "college_grad",
    "SEX AND AGE - Under 5 years": "under_5yrs"
}
```

Figure A.1: Selected Socioeconomic Fields From the FactFinder Mapped to the Test Database

Appendix B

NoSQL and Relational Databases with Geospatial Functionality

Some examples of NoSQL databases are presented below with an explanation of the tools they use to handle geospatial analysis. The relational databases considered are Postgres and MemSQL. NoSQL databases described are Redis and Neo4j.

B.1 Postgres Database with PostGIS Extension

PostgreSQL (or Postgres) is a reliable open source object-relational database management system that runs on most major operating systems, is fully ACID compliant and has native programming interfaces for most programming languages. PostGIS is an open source PostgreSQL extension that follows the Simple Features for SQL specification defined by the Open Geospatial Consortium (OGC). It adds various spatial functions and data types to PostgreSQL and makes it easier and more efficient to store and query geospatial data in a PostgreSQL database.

B.2 Redis

Redis is an open-source data structure store that operates in memory which makes it highly available. It is used as a database, cache and message broker, and supports several data structures as well as geospatial querying. Redis stores geographic positioning using Geohashes. These Geohashes, together with the geospatial commands (which include GEOADD, GEODIST, GEOHASH, GEOPOS, GEORADIUS, GEORADIUSBYMEMBER) enable Redis to implement various geospatial queries.

B.3 MemSQL and MemSQL GeoSpatial

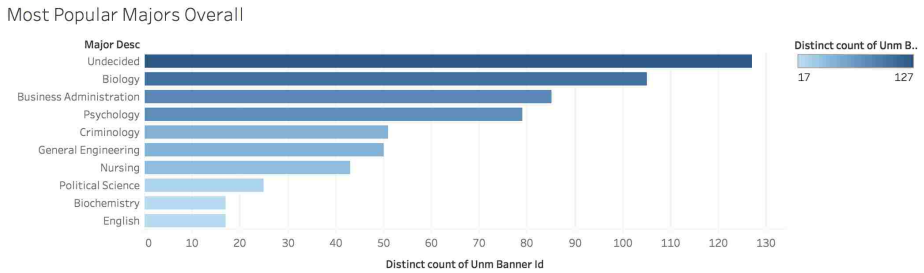
MemSQL is an in-memory database much like redis. It is however, a relational database and scales horizontally like most distributed systems. MemSQL is fast, since it stores data in-memory. It also performs well with real-time analytical and transactional processing. MemSQL supports geospatial datatypes, topological functions and and measurement functions. MemSQL geospatial is a partial implementation of the OpenGIS standard for geospatial processing. It is able to perform relational, temporal and spatial transactions and analysis at a massive scale and high performance.

B.3.1 Neo4j and Neo4j Spatial

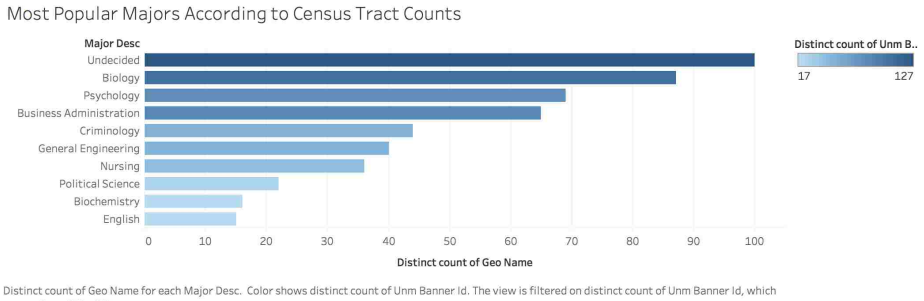
Neo4j is an ACID-compliant transactional graph database implemented in Java and Scala. It is open source and implements the Property Graph Model. Neo4j Spatial is a library that provides the tools needed to run geospatial operations on a Neo4j database.

Appendix C

Popularity of Majors at UNM



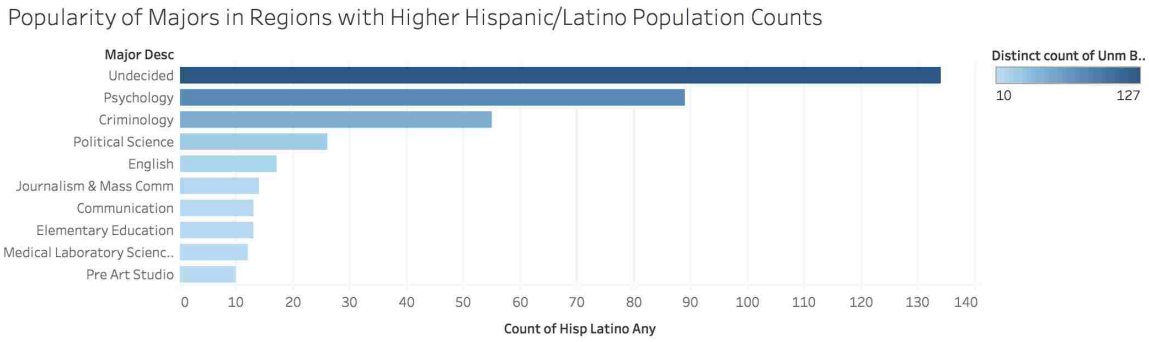
(a) Popularity of Majors at UNM for students who live in New Mexico



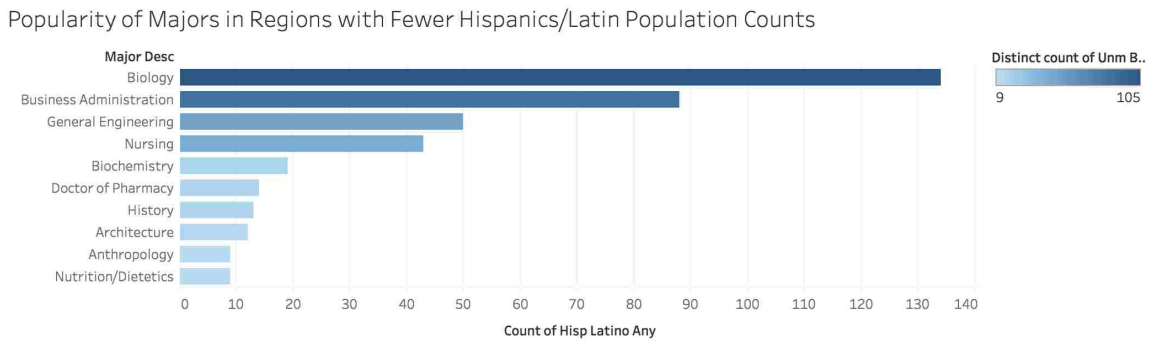
(b) Popularity of Majors According to Census Tract Distribution

Figure C.1: Overall Popularity of Majors at UNM

Appendix C. Popularity of Majors at UNM



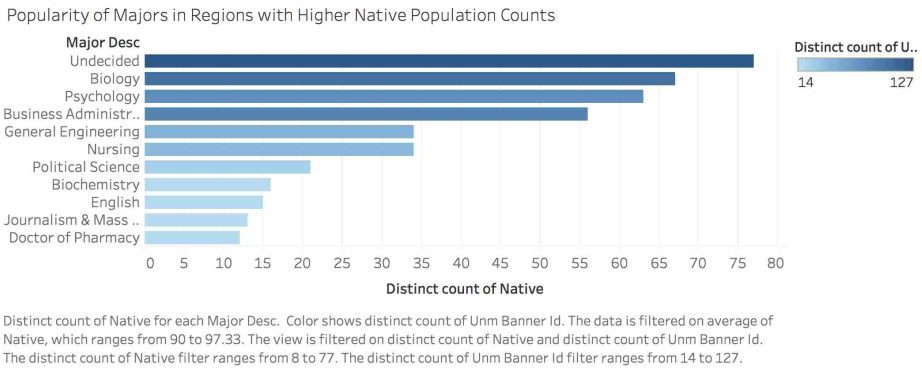
(a) Popularity of Majors in Regions with More Hispanics and Latinos



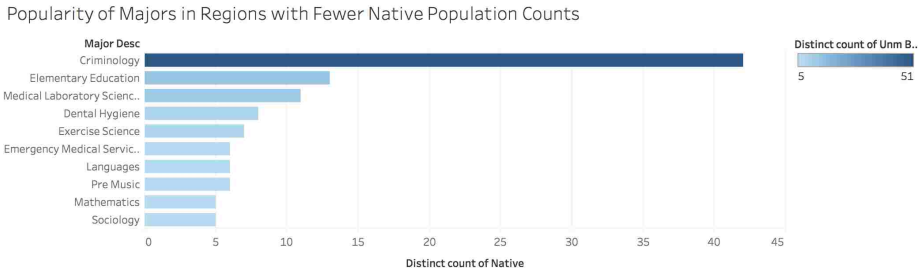
(b) Popularity of Majors in Regions with Fewer Hispanics and Latinos

Figure C.2: Popularity of Majors by Hispanic and Latino Classification

Appendix C. Popularity of Majors at UNM



(a) Popularity of Majors in Regions with More Natives



(b) Popularity of Majors in Regions with Fewer Natives

Figure C.3: Popularity of Majors by Native Classification

References

[1] A. A. for the Advancement of Science. (2017, 06) What are geospatial technologies? [Online]. Available: https://www.aaas.org/content/what-are-geospatial-technologies

[2] P. Gurin, E. Dey, S. Hurtado, and G. Gurin, "Diversity and higher education: Theory and impact on educational outcomes," *Harvard Educational Review*, vol. 72, no. 3, pp. 330–367, 2002. [Online]. Available: https://doi.org/10.17763/haer.72.3.01151786u134n051

[3] D. Nusche, "Assessment of learning outcomes in higher education: a comparative review of selected practices," *OECD Education Working Papers*, no. 15, 2008.

[4] O. of Institutional Analytics, "Official enrollment report," University of New Mexico, Tech. Rep., July 2016. [Online]. Available: https://oia.unm.edu/facts-and-figures/oer-fall-2016.pdf

[5] C. P. Gambino, "Who has a second-generation educational attainment advantage?" *SEHSD Working Paper*, vol. 2017, no. 17, 2017.

[6] A. Slim, "Curricular analytics in higher education," Ph.D. dissertation, University of New Mexico, 2016. [Online]. Available: http://digitalrepository.unm.edu/ece_etds/304

[7] T. P. D. Group, *PostGIS 2.3.3dev Manual*, svn revision (15437) ed., The PostGIS Development Group.

[8] R. Kötter, *NEUROSCIENCE DATABASES, A Practical Guide*, R. Kötter, Ed. Kluwer Academic Publishers, 2003.

[9] A. Borrmann and E. Rank, "Topological analysis of 3d building models using a spatial query language," *Advanced Engineering Informatics*, vol. 23,

no. 4, pp. 370 – 385, 2009, civil Engineering Informatics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034609000287

[10] B. Baas, "Nosql spatial: Neo4j versus postgis," Ph.D. dissertation, Delft University of Technology, 2012.

[11] Y. Lu, "Geospatial data indexing analysis and visualization via web services with autonomic resource management," Ph.D. dissertation, Florida International University, 2013.

[12] K. B. Camille L. Ryan, "Educational attainment in the united states: 2015," March 2016.

[13] J. Baker, "Bber population estimates for new mexico, 2001-2006: Origins of a growing gap with census bureau estimates." pp. 1–2, 6–7, 2007.

[14] M. M. Jack Baker, Xiaomin Ruan, "Spatial demography as a method for population estimation: Addressing census bureau under-estimation of new mexico's populations using gis technologies," 2007.

[15] K. B. Jessica Davis, "School enrollment in the united states: 2011," September 2013.

[16] N. R. Council, *Modernizing the U.S. Census*, B. Edmonston and C. Schultze, Eds. Washington, DC: The National Academies Press, 1995. [Online]. Available: https://www.nap.edu/catalog/4805/modernizing-the-us-census