Geospatial Operational Insights for National Library Board (NLB)

Project Proposal

ANLY482 – Analytics Practicum AY16/17 Term 1

Prepared by:

Team Qui Vivra Verra

LIU Bowei

PONG Chong Xin

TEO Hui Min

Supervised by:

Prof KAM Tin Seong

Associate Professor of Information Systems

Table of Contents

1. Project Overview

1.1 Background Information

The National Library Board (NLB) is a statutory board of the Ministry of Communications and Information in Singapore. Preserving a sizeable amount of title collections in the regional and public libraries which it manages, the NLB seeks to serve as a reference source for the Singapore population to connect with the precious archives of the past. Strategically scattered over the island, the NLB has up to 27 Regional and Community Libraries under its arm[1]. Beside from having a physical heritage collection, the young and the old can also access digitised materials and resources online.

1.2 Project Sponsor & Liaison Information

Our project sponsor is the National Library Board.

1.3 Project Motivation & Problem Statement

In this age of information, we see an increasing need for people and businesses to have a greater access to space and resources to further their personal and corporate needs. Hence, there is the requisite for the libraries to adequately manage this associated increasing demand. However, there exists this difficulty in measuring the operational readiness of the libraries; unlike typical corporations and organisations, the measure of public demand is not in dollars and cents.

Furthermore, there have been renovations and relocation of existing libraries and unveiling of new libraries to keep up with the times. These constant changes prompt for a reliable system to measure the effectiveness of past policies, as well as an accurate predictive model to conduct what-if analyses for future plans. A user-friendly system which displays geospatial information that can provide operational insights would thus be valuable to the NLB.

[1] See http://www.nlb.gov.sg/VisitUs.aspx for the official list of libraries managed by NLB.

1.4 Project Objectives

The main aim of the project is to provide NLB with valuable operational insights by developing a geospatial dashboard contained in a web-application, which determines the following when an existing library is relocated/removed or when a new library is added:

a. Demand capture area of libraries
b. Patronage levels of libraries
c. Associated operational-related variables e.g. subzones served, distance to the nearest transport network (MRT station and/or bus stop)

To ensure the continued sustainability of the web-application, end-users will be able to upload files of the following format to update the model:

a. .csv format
b. .xlsx format

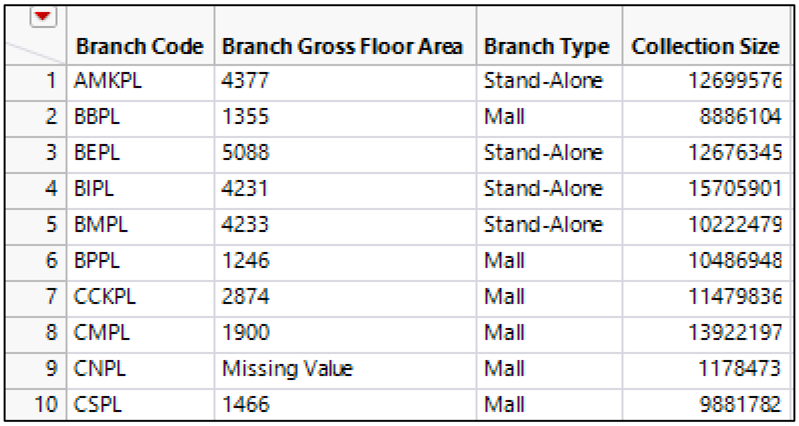
2. Data Provided

Currently, we are provided with the following datasets that give an overview of the borrowing/transaction trends in all the libraries in 2013 and 2014:

a. Collection_Dataset_FY13 and FY14.xlsx
b. Patron_Dataset_FY13.csv
c. Patron_Dataset_FY14.csv
d. Patron_Headers.csv
e. TXN_FY13.csv
f. TXN_FY14.csv
g. TXN_Headers.csv

2.1 Interpretation of Data

A sample of the *Collection_Dataset_FY13 and FY14.xlsx* dataset is as shown below:

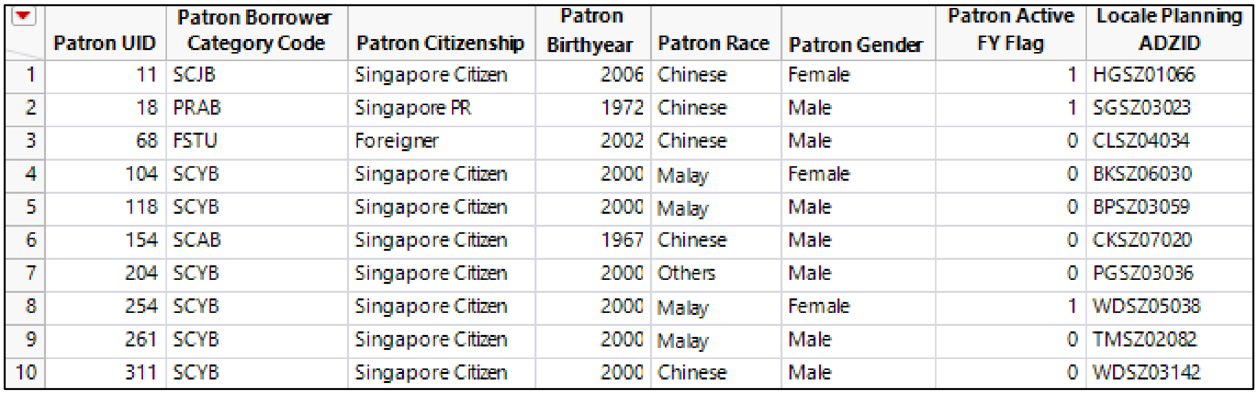


	Branch Code	Branch Gross Floor Area	Branch Type	Collection Size
1	AMKPL	4377	Stand-Alone	12699576
2	BBPL	1355	Mall	8886104
3	BEPL	5088	Stand-Alone	12676345
4	BIPL	4231	Stand-Alone	15705901
5	BMPL	4233	Stand-Alone	10222479
6	BPPL	1246	Mall	10486948
7	CCKPL	2874	Mall	11479836
8	CMPL	1900	Mall	13922197
9	CNPL	Missing Value	Mall	1178473
10	CSPL	1466	Mall	9881782

The interpretation of the *Collection_Dataset_FY13 and FY14.xlsx* is as follows:

S/N	Column Heading	Interpretation
1	Branch Code	Unique ID assigned to a library
2	Branch Gross Floor Area	Gross floor area of a specified library
3	Branch Type	A geographical classifier which takes on the value of “Stand-Alone”, “Mall”, “Regional”, or “MOLLY”
4	Collection Size	Total number of titles stored in the library

A sample of the *Patron_Dataset_FY13.csv* and *Patron_Dataset_FY14.csv* dataset is as shown below:

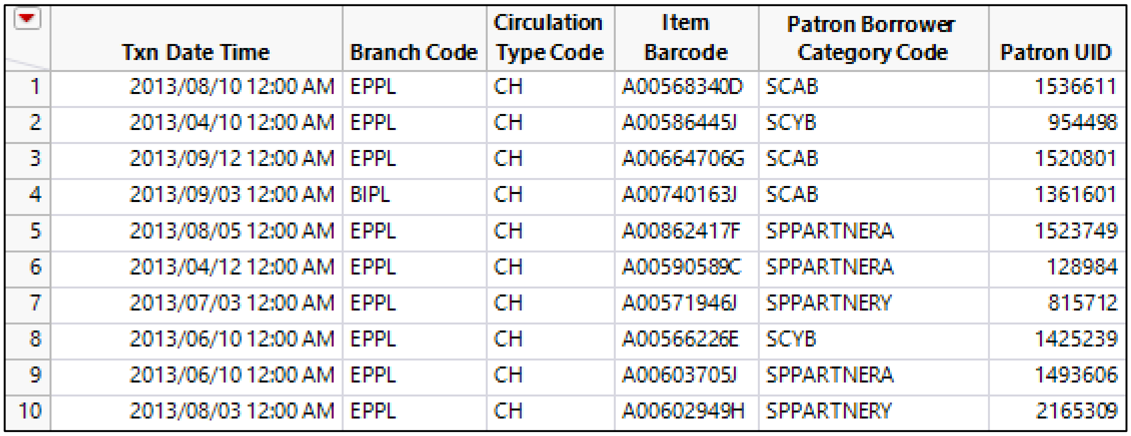


	Patron UID	Patron Borrower Category Code	Patron Citizenship	Patron Birthyear	Patron Race	Patron Gender	Patron Active FY Flag	Locale Planning ADZID
1	11	SCJB	Singapore Citizen	2006	Chinese	Female	1	HGSZ01066
2	18	PRAB	Singapore PR	1972	Chinese	Male	1	SGSZ03023
3	68	FSTU	Foreigner	2002	Chinese	Male	0	CLSZ04034
4	104	SCYB	Singapore Citizen	2000	Malay	Female	0	BKSZ06030
5	118	SCYB	Singapore Citizen	2000	Malay	Male	0	BPSZ03059
6	154	SCAB	Singapore Citizen	1967	Chinese	Male	0	CKSZ07020
7	204	SCYB	Singapore Citizen	2000	Others	Male	0	PGSZ03036
8	254	SCYB	Singapore Citizen	2000	Malay	Female	1	WDSZ05038
9	261	SCYB	Singapore Citizen	2000	Malay	Male	0	TMSZ02082
10	311	SCYB	Singapore Citizen	2000	Chinese	Male	0	WDSZ03142

The interpretation of the *Patron_Dataset_FY13.csv* and *Patron_Dataset_FY14.csv* is as follows:

S/N	Column Heading	Interpretation
1	Patron UID	Unique ID assigned to a library member
2	Patron Borrower Category Code	A classifier attached to a library member's transaction type
3	Patron Citizenship	Indicates a library member's citizenship
4	Patron Birthyear	Indicates a library member's birth year
5	Patron Race	Indicates a library member's race
6	Patron Gender	Indicates a library member's gender
7	Patron Active FY Flag	A binary variable. "1" indicates the library member is active in the FY; "0" indicates the library member is inactive in the FY.
8	Locale Planning ADZID	Indicates the geographical subzone which the library member is located in (address)

A sample of the *TXN_FY13.csv* and *TXN_FY14.csv* dataset is as shown below:



	Txn Date Time	Branch Code	Circulation Type Code	Item Barcode	Patron Borrower Category Code	Patron UID
1	2013/08/10 12:00 AM	EPPL	CH	A00568340D	SCAB	1536611
2	2013/04/10 12:00 AM	EPPL	CH	A00586445J	SCYB	954498
3	2013/09/12 12:00 AM	EPPL	CH	A00664706G	SCAB	1520801
4	2013/09/03 12:00 AM	BIPL	CH	A00740163J	SCAB	1361601
5	2013/08/05 12:00 AM	EPPL	CH	A00862417F	SPPARTNERA	1523749
6	2013/04/12 12:00 AM	EPPL	CH	A00590589C	SPPARTNERA	128984
7	2013/07/03 12:00 AM	EPPL	CH	A00571946J	SPPARTNERY	815712
8	2013/06/10 12:00 AM	EPPL	CH	A00566226E	SCYB	1425239
9	2013/06/10 12:00 AM	EPPL	CH	A00603705J	SPPARTNERA	1493606
10	2013/08/03 12:00 AM	EPPL	CH	A00602949H	SPPARTNERY	2165309

The interpretation of the *TXN_FY13.csv* and *TXN_FY14.csv* is as follows:

S/N	Column Heading	Interpretation
1	Txn Date Time	Indicates the date and time of the day during when the specified transaction took place
2	Branch Code	A unique identifier which indicates the library where the specified transaction took place

3	Circulation Type Code	A unique identifier which identifies the circulation type
4	Item Barcode	A unique identifier which indicates the item which is transacted
5	Patron Borrower Category Code	A classifier attached to a library member's transaction type
6	Patron UID	Unique ID assigned to a library member

The datasets *Patron_Headers.csv* and *TXN_Headers.csv* contain the column headings to the *Patron_Dataset_FY13.csv, Patron_Dataset_FY14* and *TXN_FY13.csv, TXN_FY14.csv* respectively.

2.2 Additional Data

The team has derived the following data from online sources (e.g. https://data.gov.sg) to complement the data provided as to ensure the completeness of the analyses to be performed. The data can be categorized into 3 categories elaborated below.

a. Facility Dataset:

   i. *Geographical location of Shopping Malls/ Plazas*

   The team recognises the positive inter-store externalities generated by the shopping malls that operate near the library (Brueckner, 2011), as more consumers visit the shopping malls, the patronage level of the nearby library will likely follow a similar increase. Hence, the presence of shopping malls/plazas near a library will contribute significantly to the attractiveness of a library.

   ii. *Geographical location of Primary Schools/ Secondary Schools/ Junior Colleges*

   A library that is located near educational institutions such as primary schools, secondary schools and junior colleges may also draw the student crowd after school hours and during the weekends. Students may also utilise the study spaces in the libraries to revise for the upcoming examinations. Hence, locating geographically nearer to an educational institution may also contribute to the attractiveness of a library.

iii. *Geographical location of Childcare Centres and Tuition Centres*

The tuition scene of Singapore has experienced a boom in recent years, as more and more parents send their children to attend additional classes after school hours (Varma, 2016). As the children wait for their tuition classes to start, and as parents wait for their kids' classes to end, a nearby library may be a go-to spot for these groups to kill some time. Hence, the team will also look at the list of all registered tuition centres in Singapore and contrast it with the locations of nearby libraries, recognising that a library is able to draw a higher patronage level with more tuition centres located nearby.

b. Transport Dataset:

   i. *Geographical location of MRT Stations (A greater weight will be assigned to MRT interchanges in the analyses)*

   ii. *Geographical location of Bus Stops & No. of Bus Services Provided*

The geographical proximity of transport systems such as the MRT and bus network cannot be neglected when estimating the attractiveness of a library. MRT stations and bus stops can be seen as network clusters of a particular subzone, where there is a high exchange of people within the areas. Furthermore, there is the greater accessibility attached to a particular library if it is located near to MRT stations and have several bus stops within walking distance. In our analyses, a greater weight will be assigned to MRT interchanges, bus interchanges, and bus stops which provide more bus services.

c. Geographical Dataset:

   i. Subzone areas of Singapore

ii. Population per subzone

iii. Land-use zoning plan for each subzone

To utilise the various datasets, the geographical location information from the Collection Dataset can be matched to subzones. The same can be done with the Patron Dataset to determine the number of NLB patrons within each subzone. The same matching process can also be applied to the Transaction Dataset to determine the number of patrons within each subzone that visited each library.

3. Methodology

3.1 Data Preparation

Further analysis of the data set can be accomplished through market segmentation. The concept of k-means clustering can be applied on the Transaction Dataset, with the clustering parameters set as: *Recency* (number of days from last transaction to end of the FY), *Frequency* (number of transactions performed within the FY) and *Monetary* (average number of books borrowed per transaction)[2]. Each patron will then be assigned to a cluster, with each cluster homogeneous within and heterogeneous across. From here, we can determine the dominant cluster of library member that each library caters to – which can provide some operational insights by understanding the demographics of the bulk of each library's patrons.

3.2 Applying the Huff's Model

An adaptation of the Huff's Model (Huff, 1964) will be applied in the analyses.
To quote a paper by Okabe & Sugihara (2012):

> “ To state a general form of the Huff model, we consider a space S (which may be a plane or a network), in which n stores are located at $p_1, \ldots, p_n$. Let a_i be the attractiveness of store i, which may be a function of its floor area, the number of

[2] Adapted from *Using datamining techniques for profiling profitable hotel customers: An application of RFM analysis* (Dursun & Caber, 2016)

items sold, its parking area and so forth; let $d(p, p_i)$ be the distance between a point p on S and the store at p_i, which may be the Euclidean distance or the shortest-path distance; and let $F(d(p, p_i))$ be a monotonically decreasing function of $d(p, p_i)$, referred to as a *distance decay function* or *distance deterrence function*. In these terms, the Huff model showing the probability of a consumer at p choosing the store at p_i is generally written as:

$$P_i(p) = \frac{a_i F(d(p, p_i))}{\sum_{k=1}^{n} a_k F(d(p, p_k))}.$$

”

Adapting the Huff's Model to the context of our project, we would consider Singapore as space S, in which n libraries are located at $p_1, \ldots, p_n$. Let a_i be the attractiveness of library *I*, which is estimated by a multinomial generalised linear regression equation, taking into account the following factors (non-exhaustive):

a. Size of the library's collection
b. Gross floor area of the library
c. Type of facility the library is located in (i.e. mall, stand-alone etc)
d. Size of facility the library is in (i.e. if the library is located in a mall, this refers to the gross floor area of the mall)
e. Number of MRT stations within a set distance (to be determined) from the library
f. Number of bus stops within a set distance (to be determined) from the library
g. Number of bus routes within a set distance (to be determined) from the library
h. Opening hours of the library
i. Number of educational institutes (i.e. primary/secondary schools, junior colleges, polytechnics, ITE, universities) within a set distance (to be determined) from the library
j. Number of other libraries (only considering the list under NLB) within a set distance from the library

Let $d(p, p_i)$ be the distance between an area (geographical subzone) p on S and the library at p_i, which may be the Euclidean distance or the shortest-path distance; and let $F(d(p, p_i))$ be a monotonically decreasing function of $d(p, p_i)$, referred to as a *distance decay function* or *distance deterrence function*. Therefore, the above-stated formula can be interpreted as the probability of a consumer at p choosing the library at p_i.

Dividing the number of patrons in each subzone at p that visited a library p_i by the total number of patrons in the subzone at p, we can obtain a probabilistic model which estimates the proportion of time that a patron from subzone p will visit library i in any given FY. Then, by substituting the known values of a_i (to be determined by the regression model) and $d(p, p_i)$ into the adapted Huff's Model, we are able to derive possible values of the power parameter ($\propto$) that govern the *distance decay function* By doing this process iteratively, we can obtain an unbiased estimate for $\propto$ that is accurate to a certain significant level.

4. Technology

For our project, we will be utilising the following technologies/tools.

4.1 JMP Pro 12

JMP Pro 12 is a tool developed by the JMP division of SAS. As the data files are too large to be opened by conventional means such as Excel and Notepad, we will be using this tool to explore the data. Market Segment Analysis will also be done using the clustering function of this application.

4.2 Leaflet

Leaflet.js is an open source javascript library for interactive maps. This tool will be used to create a visualization page for the users where a map of Singapore, as well as point symbols representing various facilities will be displayed. The user can select the attribute to be considered for computing the attractiveness index by selecting or deselecting facility layers as well as varying buffer radius. This tool is selected as it provides a range of interactive maps and is easy to implement. It supports various plugins to extend its functionality.

4.3 JavaScript

JavaScript is a coding language for the web. We will be using JavaScript for most of the application's user interfaces as it allows the implementation of various libraries to support user's interactions and improve visualisation.

4.4 Turf.js

Turf.js was mainly used for spatial analysis. It provides the functionality to analyse, aggregate and transform data into GeoJSON.

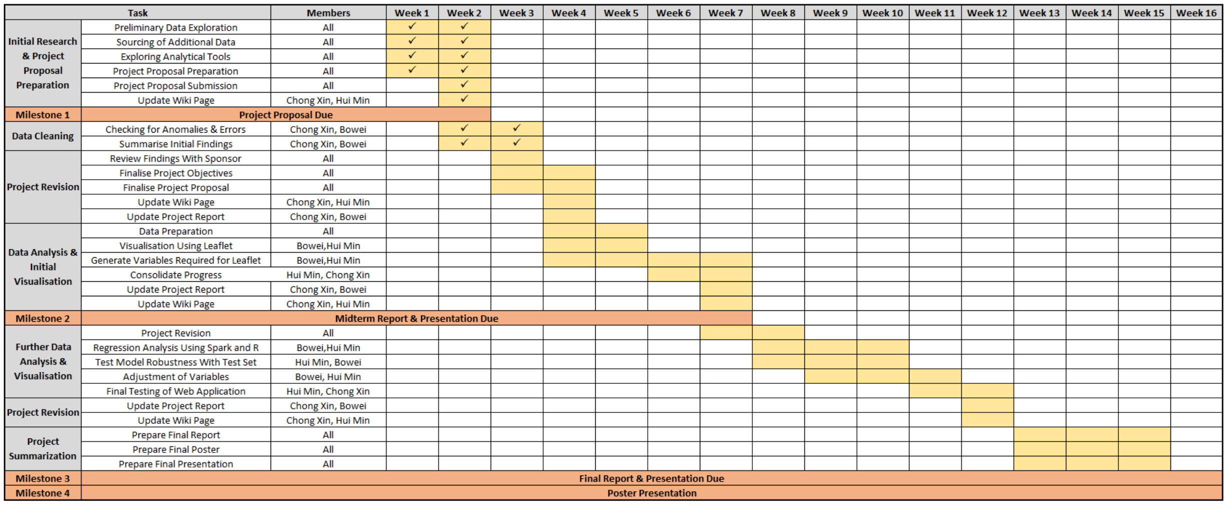
4.5 SQLite and SpatiaLite

SQLite and SpatiaLite extension will be used as a database to store the geospatial data uploaded by the user. SpatiaLite will then be used to query from the database variables needed for the huff's model.

4.6 Apache Spark

We will be using Apache Spark's Machine Learning Library for performing regression analysis on the huff's model.

5. Timeline & Schedule



Task		Members	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15	Week 16
Initial Research & Project Proposal Preparation	Preliminary Data Exploration	All	✓	✓														
	Sourcing of Additional Data	All	✓	✓														
	Exploring Analytical Tools	All	✓	✓														
	Project Proposal Preparation	All	✓	✓														
	Project Proposal Submission	All		✓														
	Update Wiki Page	Chong Xin, Hui Min		✓														
Milestone 1	Project Proposal Due																	
Data Cleaning	Checking for Anomalies & Errors	Chong Xin, Bowei		✓	✓													
	Summarise Initial Findings	Chong Xin, Bowei		✓	✓													
Project Revision	Review Findings With Sponsor	All			🟨													
	Finalise Project Objectives	All			🟨	🟨												
	Finalise Project Proposal	All			🟨	🟨												
	Update Wiki Page	Chong Xin, Hui Min				🟨												
	Update Project Report	Chong Xin, Bowei				🟨												
Data Analysis & Initial Visualisation	Data Preparation	All				🟨	🟨											
	Visualisation Using Leaflet	Bowei,Hui Min				🟨	🟨											
	Generate Variables Required for Leaflet	Bowei,Hui Min				🟨	🟨	🟨	🟨									
	Consolidate Progress	Hui Min, Chong Xin						🟨	🟨									
	Update Project Report	Chong Xin, Bowei							🟨									
	Update Wiki Page	Chong Xin, Hui Min							🟨									
Milestone 2	Midterm Report & Presentation Due																	
Further Data Analysis & Visualisation	Project Revision	All							🟨	🟨								
	Regression Analysis Using Spark and R	Bowei,Hui Min								🟨	🟨	🟨						
	Test Model Robustness With Test Set	Hui Min, Bowei								🟨	🟨	🟨						
	Adjustment of Variables	Bowei, Hui Min									🟨	🟨	🟨					
	Final Testing of Web Application	Hui Min, Chong Xin											🟨	🟨				
Project Revision	Update Project Report	Chong Xin, Bowei												🟨				
	Update Wiki Page	Chong Xin, Hui Min												🟨				
Project Summarization	Prepare Final Report	All													🟨	🟨	🟨	
	Prepare Final Poster	All													🟨	🟨	🟨	
	Prepare Final Presentation	All													🟨	🟨	🟨	
Milestone 3	Final Report & Presentation Due																	
Milestone 4	Poster Presentation																	

6. Risks & Limitations

Risks & Limitations	Mitigation Strategy
Lack of experience with analytical tools (i.e. Apache Spark, SpatiaLite, JMP Pro 12)	Explore and familiarise with the analytical tools prior to using them to perform the actual analyses. Use the week before iteration one as the study week to learn the necessary skills required to wield the tools.
Changes in work schedule due to unexpected events; delay of first release and other milestones	Raise awareness about the change and re-look the work breakdown structure and tasks allocation.
Presence of other projects that will potentially hinder the project progress	Priority of this project is emphasized. The introduction of a buffer week will also help reduce the associated impact.

7. References

Brueckner, J. (2011). *Lectures on Urban Economics*. The MIT Press.

Varma, A. (2016, May 15). More primary and secondary school students are getting private tuition years in advance of their grade in school. Retrieved August 25, 2016, from *http://www.straitstimes.com/lifestyle/more-primary-and-secondary-school-students-are-getting-private-tuition-years-in-advance-of*

Dursun, & Caber. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives, 18*, 153-160.

Huff, D. (1964). Defining and Estimating a Trading Area. *Journal of Marketing, 28*(3), 34-38.

Okabe, Atsuyuki, & Sugihara, Kokichi. (2012). Network Huff Model. In *Statistics in Practice* (pp. 213-230). Chichester, UK: John Wiley & Sons.