****

**Geospatial Operational Insights**

**for National Library Board (NLB)**

**Project Initiation Proposal**

**ANLY482 – Analytics Practicum AY16/17 Term 2**

**Prepared by:**

*Team Bookworms*

GAURI BHATNAGAR

THAVANESAN S/O SIVANANTHAN

WANG TIANTONG

**Supervised by:**

Prof KAM Tin Seong

*Associate Professor of Information Systems*

Table of Contents

[1. Project Overview 3](#_Toc472264769)

 [1.1 Background Information 3](#_Toc472264770)

 [1.2 Project Sponsor & Liaison Information 3](#_Toc472264771)

 [1.3 Project Motivation & Problem Statement 3](#_Toc472264772)

 [1.4 Project Objective 4](#_Toc472264773)

 [1.5 Desired Outcome 4](#_Toc472264774)

2. [Data 4](#_Toc472264775)

 [2.1 Raw Data from Sponsor 4](#_Toc472264776)

 [2.1.1 ‘Collection\_Dataset\_FY13 and FY14.xlsx’ 4](#_Toc472264777)

 [2.1.2 ‘Patron\_Headers.csv’ 5](#_Toc472264778)

 [2.1.3 ‘TXN\_FY13.csv & TXN\_FY14.csv’ 6](#_Toc472264779)

 [2.2 Additional Data 6](#_Toc472264780)

 [2.2.1 Surrounding Facility Dataset: 7](#_Toc472264781)

 [2.2.2 Transportation Accessibility Dataset: 7](#_Toc472264782)

 [2.2.3 Geographical Dataset: 7](#_Toc472264783)

 [2.2.4 Field Research Data – Internal Facilities and Patron Preference 8](#_Toc472264784)

[3. Methodology 8](#_Toc472264785)

 [3.1 Data Pre-Processing 8](#_Toc472264786)

 [3.2 Old Model Calibration 8](#_Toc472264787)

 [3.3 Attractiveness Measurements Determination 9](#_Toc472264788)

 [3.4 Clustering Analysis 9](#_Toc472264789)

 [3.5 Adaptation of Huff’s Model and Multiplicative Competitive Interaction Model 9](#_Toc472264790)

[4. Technology 10](#_Toc472264791)

 [3.1 For Data Preparation & Model Calibration 11](#_Toc472264792)

 [3.1.1 R Package - Multiplicative Competitive Interaction (MCI) Model 11](#_Toc472264793)

 [3.1.2 R Programming Language 11](#_Toc472264794)

 [3.1.3 JMP Pro 13 11](#_Toc472264795)

 [3.2 For Dashboard Visualization 11](#_Toc472264796)

 [3.2.1 R Shiny & R Markdown 11](#_Toc472264797)

 [3.2.2 Javascript 11](#_Toc472264798)

 [3.2.3 Leaflet 12](#_Toc472264799)

[5. Timeline & Milestones 12](#_Toc472264800)

[6. Risks & Limitations 12](#_Toc472264801)

[7. References 13](#_Toc472264802)

# Project Overview

## 1.1 Background Information

The National Library Board (NLB) is a statutory board of the Ministry of Communications and Information in Singapore, tasked to manage the public libraries and lead them into the information age where non-print resources are making their mark. With a substantial amount of title collections in the regional and public libraries which it manages, the NLB aims to aid as a point source for Singaporeans to connect with the precious archives of the past. The NLB has managed to disperse its 27 Regional and Community Libraries strategically across Singapore.

## 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 increase in necessity for people and businesses to have a greater access to space and resources to further their personal and corporate needs. Libraries in Singapore are no exception to this trend.

To meet this necessity, NLB may not only have to find ideal locations within districts for additional libraries, but it also must manage their resources wisely to meet public demand with respect to each location. Currently there exists this difficulty in measuring the operational readiness of the libraries given that the measure of public demand is not in dollars and cents; unlike that of typical corporations and organizations.

Constant change management in the location and resources within the libraries have led to the dire need of a reliable and standardized method which can measure the effectiveness of past policies, as well as an accurate predictive model to conduct what-if analyses for future. Thus, a user-friendly system which displays geospatial information based on significant data evidence would provide operational insights valuable to the NLB.

Besides taking geospatial information into consideration, NLB is also motivated to leverage on an effective model (which uses relevant metrics regarding book collections and user transaction information) to accurately measure the attractiveness of existing and potential libraries.

## 1.4 Project Objective

Develop a geospatial dashboard that would facilitate the NLB’s evaluation of different potential library siting locations through better prediction of the catchment and number of patrons it can potentially attract.

Assess earlier models which have been used to evaluate the attractiveness of libraries and refine them to increase the accuracy and reliability of results.

## 1.5 Desired Outcome

The geospatial dashboard will potentially contribute to:

* the planning/promotion of content and services that are localised to the specific population to be served
* the evaluation of the attractiveness of existing and future NLB library locations

# Data

## 2.1 Raw Data from Sponsor

Currently we have received the raw data provided by our sponsor including the following files:

* Collection\_Dataset\_FY13 and FY14.xlsx
* Patron\_Headers.csv
* TXN\_FY13.csv & TXN\_FY14.csv

### 2.1.1 ‘Collection\_Dataset\_FY13 and FY14.xlsx’

#### Metadata

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Attribute Type** | **Attribute Meaning** |
| Branch Code | Text | Unique ID assigned to a library |
| Branch Gross Floor Area | Number (with “missing value”) | Gross floor area of a specified library |
| Branch Type | Text | A geographical classifier (“Mall”/ “Stand-Alone”/ “MOLLY”/ “Regional”) |
| Collection Size | Number | Total number of titles stored in the library |

#### Sample Data



### 2.1.2 ‘Patron\_Headers.csv’

This is the file that stores headers of *‘Patron\_Dataset\_FY13.csv and Patron\_Dataset\_FY14.csv’* which will be collected from our sponsor in the later phase. Regarding patron’s data, our team will also look at the post-geocoding data provided by our sponsor in the later phase, which includes the X-Y coordinates of the customers visiting each library. With the data, we will build a more precise and accurate before/after model and dashboard at the HDB level for the evaluation of potential location choices of the libraries.

#### Metadata

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Attribute Type** | **Attribute Meaning** |
| Patron ID | Number | Unique ID assigned to a library member  |
| Patron Borrower Category Code | Text | A classifier attached to a library member’s transaction type  |
| Patron Citizenship | Text | Indicates a library member’s citizenship  |
| Patron Birthyear | Number | Indicates a library member’s birth year  |
| Patron Race | Text | Indicates a library member’s race  |
| Patron Gender | Text | Indicates a library member’s gender  |
| Patron Active FY Flag  | Number (0/1) | A binary variable. “1” indicates the library member is active in the FY; “0” indicates the library member is inactive in the FY.  |
| Locale Planning ADZID  | Text (Combination of letters and numbers) | Indicates the geographical subzone which the library member is in (address)  |

### 2.1.3 ‘TXN\_FY13.csv & TXN\_FY14.csv’

#### Metadata

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Attribute Type** | **Attribute Meaning** |
| Txn Date Time  | Date time (m/d/yyyy h:mm) | Indicates the date and time of the day during when the specified transaction took place  |
| Branch Code  | Text | A unique identifier which indicates the library where the specified transaction took place  |
| Circulation Type Code  | Text | A unique identifier which identifies the circulation type  |
| Item Barcode  | Text(Letters & Numbers) | A unique identifier which indicates the item which is transacted  |
| Patron Borrower Category Code  | Text | A classifier attached to a library member’s transaction type  |
| Patron UID  | Number | Unique ID assigned to a library member  |

#### Sample Data



## 2.2 Additional Data

Based on the senior’s group’s work, our team will conduct the model building and calibration using geospatial data derived from online sources (e.g. https://data.gov.sg) to evaluate and update the corresponding attractiveness measurements. Our team will also do field research to collect relevant data on library layout design, collection availability, library facilities and patron preferences in person. At the first stage, the data can be categorized into 4 categories elaborated below.

### 2.2.1 Surrounding Facility Dataset:

#### Geographical location of Shopping Malls/ Plazas

As indicated in the senior’s group’s report, there is *“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*.” Therefore, our team will keep studying the significant effect on the patronage of the libraries from the distribution of various shopping malls.

#### Geographical location of Primary Schools/ Secondary Schools/ Junior Colleges

As one of the largest groups visiting libraries, students are nonnegligible given that they are likely to spend time in the libraries after school hours and during examination period. Hence, our team will also have a deep look at the impact on the patronage of the libraries based on the location distribution of nearby educational institutions (primary schools, secondary school, junior colleges) using the data derived online.

### 2.2.2 Transportation Accessibility Dataset:

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

#### Geographical location of Bus Stops & No. of Bus Services Provided

#### In order to evaluate the likelihood for a patron to visit a library, the accessibility of transportation also plays an important role. With an easily accessed public transport network connected to a library, there will be less hindrance and thus a higher probability for a patron to visit the library. To analyze more deeply, the impact of public transportation may also vary between different neighborhoods where people are of different social and financial levels. Therefore, our team will embrace the available transportation dataset (MRT and bus stops) in our model with weight assigned to better measure and predict the attractiveness of the libraries.

### 2.2.3 Geographical Dataset:

#### building within costaloutline.shp



As mentioned above, although the subzone clustering analysis conducted by the senior’s team returned a relatively executable model, our team aims to build up on the next level and present a more precise and accurate analysis. In terms of the geographical dataset, subzones no longer meet our demand due to the wide coverage of each subzone and the inequality analysis on patrons from different parts within the same subzone. Therefore, our team will utilize the geospatial data at HDB level (after transformation) and link it to the post-geocoding patron’s data so as to better analyze the patronage of each library.

### 2.2.4 Field Research Data – Internal Facilities and Patron Preference

In real-world analytics, complete data may not be available for our target features so we need to do some filed research to collect useful data ourselves. In this project, our team believes that beside surrounding facilities and public transportations, the libraries’ layout design and their internal facilities are also part of their attractiveness. Based on our own research, patrons may prefer one library just due to its availability of some specific types of facilities, like quite areas and large study space etc. What’s more, in addition to the collection size, the collection categories and the accessibility of the collections also contribute to the attractiveness based on patrons’ preferences. Therefore, our team is planning to conduct more and different field researches and interviews to collected relevant data regarding this aspect.

# Methodology

## 3.1 Data Pre-Processing

Before building model and conducting analysis, well-preprocessed data is the foundation of a successful analytics. In our project, due to the different resources of the data we have, the format and the quality of the data also vary. Therefore, prior to the model planning phase, our team will clean the data and deal with problems such as missing value, inconsistent format and mismatch data type etc.

## 3.2 Old Model Calibration

The special point about our project is that it builds up on our seniors’ project and seeks a better solution to cater to our sponsor’s requirements and needs. Therefore, before coming up with our own solution, our team will test on the existing dashboard and understand the advantages and disadvantages of it. Further, we will do model calibration using our data to examine the accuracy and feasibility of the old model so that we can better understand what upgrades are needed and where we should start build on.

## 3.3 Attractiveness Measurements Determination

This phase is one of the highlights of out project. After data collection, data interpretation and old model calibration, our team will have a better understanding of the business domain and next come up with more accurate and interesting factors to evaluate the attractiveness of the libraries. In addition to the original factors (public transportation, shopping mall and education institutions) considered by our seniors, our team will also analyze the internal facilities of the libraries, collection types and availabilities and frequent patrons’ preferences visiting each library and apply factors with significant impact in our new model.

## 3.4 Clustering Analysis

Bearing in mind that there are patterns behind frequent patrons visiting the libraries, our team will conduct clustering analysis focusing on the patrons frequently visiting each library. Each single library will be the center of each cluster and each patron will be assigned to the cluster that s/he visited the most frequently. By doing clustering analysis, we hope to find patterns (positive/negative relationships and co-relations between different factors and the target – patronage) behind the data to better serve our model building and attractiveness measurements selection.

## 3.5 Adaptation of Huff’s Model and Multiplicative Competitive Interaction Model

The model building part is another highlight of our project which will combine both Huff’s Model and Multiplicative Competitive Interaction Model.

As one of the probabilistic market area models, the Huff’s Model (Huff 1962, 1963, 1964) is the most popular gravity-based spatial interaction model for retailing and services. Based on the idea that “customer decisions are not deterministic but probabilistic”, this model enables us to calculate these probabilities, and in our case, the probabilities (Pij) for a patron (i) of a set of features to visit a specific library (j) where 0 < Pij < 1. The Huff’s Model is based on a utility function with two explanatory variables – “attractiveness” and “hindrance” of the location which are weighted by an exponent: Uij = A γ j d −λ ij.

* A*j* is a measure of attractiveness of store j, such as square footage
* D*ij*is the distance from *i* to *j*
* is an attractiveness parameter estimated from empirical observations

* is the distance decay parameter estimated from empirical observations

* *n* is the total number of stores including store*j*.

In our project, we are going to evaluate various measures of attractiveness that may be built in the Huff’s model to calculate the probabilities, including but not limited to the below factors:

1. Size of the library’s collection
2. Gross floor area of the library
3. Accessibility of surrounding facilities of the library
4. Size of surrounding facilities of the library
5. Number of points of public transportation (MRT, bus stops) within a set distance (to be determined) from the library
6. Number of educational institutes (i.e. primary/secondary schools, junior colleges, polytechnics, ITE, universities) within a set distance (to be determined) from the library
7. Opening hours of the library
8. Availability of various collections in the library
9. Number of other NLB libraries within a set distance from the library
10. Accessibility of internal facilities in the library

Another model we are going to apply in our project is the Multiplicative Competitive Interaction (MCI) Model (Nakanishi/Cooper 1974). It is an *“econometric model for analyzing market shares in a competitive environment where the market is divided in i submarkets (e.g. groups of customers, time periods or geographical regions) and served by j suppliers (e.g. firms, brands or locations)”*. Resulting market share of the suppliers (libraries in our case), this model also analyzes the attraction/utility of the alternatives in the submarket.
Different from the senior group’s model, MCI model is nonlinear but can be transformed via Ordinary Least Square (OLS) regression using the multi-step log-centering transformation and our team will also re-arrange the raw data in an interaction matrix to fix into the model.

In terms of the implementation of the two models, our team will make use of the available MCI package provided by R which includes functions enabling fitting the MCI model, MCI shares simulations, the log-centering transformation of MCI datasets and creation of interaction matrices. The package also includes nonlinear optimization algorithms for the application of Huff’s Model.

# Technology

For this project, we are planning to customize and partially implement the following tools and techniques.

## For Data Preparation & Model Calibration

### 3.1.1 R Package - Multiplicative Competitive Interaction (MCI) Model

Packages are collections of R functions, data, and compiled code in a well-defined format. Market area models are used to analyze and predict store choices and market areas concerning retail and service locations. This package implements two market area models (Huff Model, Multiplicative Competitive Interaction Model) into R, while the emphases lie on 1.) fitting these models based on empirical data via OLS regression and nonlinear techniques and 2.) data preparation and processing (esp. interaction matrices and data preparation for the MCI Model).

### 3.1.2 R Programming Language

**R** is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical softwareand data analysis. In our project, R programming language will be utilized to code back-end analytics models and front-end user interface.

### 3.1.3 JMP Pro 13

JMP Pro 13 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.

## For Dashboard Visualization

### 3.2.1 R Shiny & R Markdown

Shiny by RStudio is a web application framework for R and R Markdown from RStudio is a collection of interactive documents for R, both of which are planned to be used in our project for a more interactive and user-friendly interface.

### 3.2.2 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 visualization.

### 3.2.3 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.

# Timeline & Milestones



# Risks & Limitations

|  |  |
| --- | --- |
| **Risks & Limitations** | **Mitigation Strategy** |
| Incorrect application of models, analytical and visualization tools (i.e. Huff’s Model, MCI R, R Shiny, JMP Pro 13) due to lack of experience | Explore and familiarize with the models and tools prior to actual implementation. Also looking at how these models have been utilized in previous projects in order to identify incremental changes to be made. |
| Inability to meet deadlines due to changes in duration required to complete tasks | Re-evaluate project breakdown and task allocation as set by the timeline upon recognizing the changes  |
| Difficulty in managing responsibilities of practicum along with other university-related projects  | Setting compulsory sessions during the week which is conducive for all members to meet and work on the practicum. |

# References

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

CRAN - Package MCI. (2017). Cran.r-project.org. Retrieved 6 January 2017, from <https://cran.r-project.org/web/packages/MCI/index.html>

Huff, D. L. (1962): “Determination of Intra-Urban Retail Trade Areas”. Los Angeles : University of California.

Huff, D. L. (1963): “A Probabilistic Analysis of Shopping Center Trade Areas”. In: Land Economics, 39, 1, p. 81-90.

Huff, D. L. (1964): “Defining and Estimating a Trading Area”. In: Journal of Marketing, 28, 4, p. 34-38.

Robert I., K. (2017). Quick-R: R Packages. Statmethods.net. Retrieved 6 January 2017, from <http://www.statmethods.net/interface/packages.html>

*R (programming language)*. (2017). *En.wikipedia.org*. Retrieved 6 January 2017, from [https://en.wikipedia.org/wiki/R\_(programming\_language)](https://en.wikipedia.org/wiki/R_%28programming_language%29)

*Shiny*. (2017). *Shiny.rstudio.com*. Retrieved 6 January 2017, from <https://shiny.rstudio.com/>

Singapore Management University, (2016). *Geospatial Operational Insights for National Library Board (NLB) Project Proposal*. [online] Singapore Management University. Available at: https://wiki.smu.edu.sg/ANLY482/img\_auth.php/2/22/ANLY482\_NLB\_Proposal\_Qui\_Vivra\_Verra\_.pdf [Accessed 6 Jan. 2017].

Wieland, T. (2016, September 1). *Package ‘MCI’.* Retrieved January 6, 2017, from https://cran.r-project.org/web/packages/MCI/MCI.pdf