Spatial-Temporal Analytics with Students Data to recommend optimum regions to stay

By

ARUN KUMAR BALASUBRAMANIAN (A0163264H)
DEVI VIJAYAKUMAR (A0163403R)
RAGHU ADITYA (A0163260N)
SHARVINA PAWASKAR (A0163302W)
LI MEIYAO (A0163379U)
Spatial-Temporal Analytics with Students Data to recommend convenient regions for stay in Singapore

Singapore is one of the most expensive cities to live. Given its excessive cost of living, it is very tough on the international students who want to live their time at Singapore with most convenience and least expenditure. Thus, we performed the analysis on spatial data with Students’ movement data points given to suggest convenient areas for a student to stay in Singapore based on facilities available nearby like Hawker Centres, Cycling Path and MRT stations.

Objectives & Motivation:

Objective is to find a convenient place for students to stay based on data collected in Singapore.

To any international student coming into Singapore needs robust information about convenient places to rent. To proceed with this we analysed student’s movement data which was collected by students of NUS belonging to ISS school, exploratory spatial data analysis was done on that to find the pattern and insights. We considered that as a student basic amenities would be economical stay, cycling path, library, MRT closeness, parks, hawker centre.

Input:

OpenPath data with student’s personal location information was collected for the month of April. Data Cleaning was done to refine the data. As the data was stored in multiple mobile devices the date and time formats were inconsistent, thus, all the date and time field values were converted to a single standard format. To do more analysis, separate columns were created for date and time. The outliers were treated using R and the data points outside Singapore were ignored. Modified Name and MailID columns to obtain missing details using EXCEL.

Projection:

With Projection and Transformation tool we converted the coordinate notation from Geographic coordinate system to Projected Coordinate system (SVY21_Singapore_TM). The X, Y (Latitude, Longitude) is given as Input Coordinate System parameter for the selected Open Path data. The unique Object ID(OID), Latitude and Longitude (DDLat, DDLon) values were automatically calculated into two separate fields. This allows any new data layers to adopt to the coordinate system of the first layer and it also facilitates exploration and mapping new layers with more accuracy.

Field Derivation:

Using Field Calculator, new fields were derived to gain insights into Students data points. Students data is categorized using the field calculator of ArcMap to derive details like day of the week or part of the day when data is collected to do analysis further.

- Day finder: Categories: ISS (Day of class), Sunday, Weekday
- Part of Day: - Morning (5am to 12pm), Afternoon (12pm to 5pm), Evening (5pm to 10pm), Night (10pm to 12am), Late-night (12am to 5am)
- Mood Detector: Weekend and Weekday
Exploratory Analysis on Students Data (with derived fields: Students Data)

Late Night distribution of Students Data:

Below map tells the students data points that are collected only during late night (12:00 am -4:00 am). since the spatial data is collected late night we assume the students reside at home mostly and is believed to be their place of stay. This is again an assumption based on time and some data points can be inaccurate to the assumption.

Projected Clustered Area of Students Data:

Open path data points were concentrated more at the NUS areas as the data points were collected during which students had more frequency travelled to university for the class and assignment related work.
Mean Centre Analysis:

Mean centre represents the geographic centre for the field attributes. And here we identified the mean centre for the derived field to understand centroid of the data points representing each new derived column.

Mean Centre for DayFinder:

The above centres show that the data during ISS (i.e. Tuesday and Saturday) day is more towards university and weekday it is moving towards the city indicating students are spread away from university on other days.

Mean Centre for Part of day

Part of the day indicates that students are near the university during the evening time, from which it can be inferred that students are attending the class or studying post the class.
Mean Centre for “Mood Detector”

The mood detector, a field created to observe the mean centre according to the corporate working days, it is observed that mean centre for the weekends is more near to university as there were classes on Saturday and weekday is far more towards city indicating students are at home or travelling.

Data Visualisation -Mood Detector

Other than NUS, the weekend data points are more concentrated towards ISS due to Saturday class and towards Little India, Vivocity, Serangoon and inclined towards the MRT lines. But during the weekends, students tend to travel less and stay near the university.
Data Visualisation- Part of Day

The data points are observed based on the time of the day. Except Late-night points all data points are concentrated towards Chou Chu Knag, near to NUS and scattered across other areas.

Data points - DayFinder

Analysis of data points across NUS and nearby areas. The further variation of points is due to people who stay far from University and need to travel most of the days for class and assignment.
Directional Distribution:

Directional distribution analysis for the open path data was carried out to find along which direction was the data inclined. The standard deviation was taken as 2 so that we get a confidence interval of 95% with the newly derived fields.

DD on Mood Detector:

The ring for the weekday is more spread than weekend indicating that students are travelling more on weekday.

Day finder:

Sunday is inclined more towards north, away from MRT, it can be taken that students don’t travel much on Sunday as transportation is not available on Sunday at university or it may be skewed because of less data points on Sunday.
Spatial-Temporal Analytics with Students Data

Part of Day:

The directional distribution for evening is less spread as the students tend to travel less and mostly stay at the place maybe continuing their work or because of less data points collected during that period.

Hot-Spot Analysis (Getis-Ord) on Students’ Data Points:

The hot-spots are area of NUS and regions surrounding NUS primarily because it is students’ data.
Cluster Analysis (Anselin Local Morons I) on Students Data:

The high-density cluster is formed near the ISS area.

From the above exploration, it is evident that most student stay at university, travel to residence and they use MRT most of the time. As students would prefer to stay near university, place near to MRT station and have the potential to use Cycling path, our aim is to find and suggest better zones for student life. Thus, we need to add these layers to the student data and carry out analysis.

Process:

To get more insights into suitable dwelling, we tried to separately analyse various datasets such as HDB dwelling population data, Hawker centre location, MRT station and cycling paths.

Analysis on HDB Singapore Dwelling Data:

The Dwelling data is exported from www.data.sg.gov site which has all the dwelling zone in Singapore for 2016. The Per person field for each dwelling zone was calculated based on (Shape Area/ Total Population) to normalize the shape area.

Getis-Ord Gi*:

This Hot spot analysis reveals zones according to the population density staying in HDBs. The darker red zone are areas with more population (with 99% confidence) and the cold spots represents less population density residing in HDB.
Anselin Local Morons I - HDB Clusters:

The high-density clusters of HDB population are near areas like Bedok, Tampines, Jurong West, Woolands and Sengang. The low-density clusters of HDB population are CPD areas of Singapore with less housing opportunities.

Geographically Weighted Regression:

To find insights from student data using HDB and population data layer, geographically weighted regression is performed using variables as shown below with an assumption that data points having timestamp of late night represents student’s home. Spatial join was done on hawker centre, dwelling and open path student data layers to obtain final GWR model.

Explanatory Variables:

The count_ variable is the number of student data points per polygon, Count_1 is the number of Hawker centres per polygon, SHAPE_Area of the dwelling layer and HDB the total count of HDB per person.

The GWR results shows that model can perform with moderate accuracy having adjusted R2 of 0.31.

Spatial Autocorrelation (Moran's I) tool on the regression residuals was run to ensure that the model residuals are spatially random. Statistically significant clustering of high and/or low residuals (model under- and overpredictions) indicates that our GWR model is not accurate enough to predict.
The below map indicates regions with localised R square to find HDB population using students data, to build for students affordable HDBs. The Dark Red regions are the places where the model predicts with higher accuracy and the blue regions are the place where its prediction is poor.
As the model residuals are not random based on the spatial autocorrelation (Morons I), it cannot be used for prediction purposes. This model was just build to study the insights of students data with other layer of data. 

Thus, to find convenient places for students, the places were ranked using student data points and factors like MRT, cycling path and hawker centers availability. 

**Analysis on Hawker Centres :**

Identify Hawker Centre with buffer of 50m for the entire region of Singapore to understand the distribution of affordable food centres for students. We conducted spot analysis, hotspot and cluster analysis on hawker centre data. And they reveal nearly the same patterns. As shown below data points are cluttered more near Clarke Quay, River Valley, Little India etc. The hot spots for Hawker centre count are identified near little india, Bedok, Marina Bay sands, Orchard and near Boat Quay. The cold clusters are some part of Punggol, Jurong Park, Simpang and Singapore Zoo.

**Hawker Centre- Spot Analysis:**
 Hop-Spot Analysis:

Cluster Analysis:
Analysis on MRT station location:
The distribution of MRT station across Singapore for Orange, Green, Purple and Red lines.

Relating MRT and Student data points:
The data points near the MRT stations with a buffer area of 75m and students’ data points near NUS with a buffer area of 5m are extracted separately.

The NUS region is considered as single polygon and only data points outside NUS area is considered.

Resulting layer of students’ data points after buffering is spatially joined on MRT data layer to understand the travelling nature of students.

Nearly 22% of data points near NUS campus and 33% along the MRT line. Here, MRT lines are extensively used by students.
Cycling Path Road Map of Singapore:

Using cycles can be economical mode of transportation for students, cycling paths around Singapore were analysed.

Park connector loop with buffer 50m:

The existing park connector loop for Singapore was taken considering that students can cycle on the park connector loop as well.

Cycling path with buffer 50m

The existing cycling path for Singapore was taken with a buffer of 50m considering students have a tolerance level of 50m.
The Final Layer on Cycling:

Both cycling path and park connector loop intersecting process was done on both cycling and park connector loop with a tolerance level of 5 meters, if the students will have a tolerance level of 5m to connect to park connector and cycling path. Then Union operation is performed with both cycling and park connector to get one single overview of the complete cycling path of Singapore with the assumption that people will cycle on park connector as well.

Converting to Raster:

All the three layers’ hawker centre, MRT stations and Cycling path were converted to raster data, so that we can easily process the continuous data and to store multiple layer of theme.

Converting to Raster using Calculate Field (Derived field):
Spatial-Temporal Analytics with Students Data

Raster Conversion for MRT data layer:

[Image of raster conversion for MRT data layer]

Raster representation for Cycling Data:

[Image of raster representation for Cycling Data]
Output:

Ranking zones in Singapore:

Based on Hawker Centres:

To get a density of hawker centres in each we calculated a new field which was used for ranking:

ValueField =

If (Student Count > 0 AND HawkerCenterCount > 0 AND Population > 0)

Then {Total/HawkerCenterCount}

Else

{99999999}

The above field derivation just indicates that how many hawker centre is needed to cater people. For example:

If total population is 1000 and hawker centre is 5, ValueField is 1000/5 = 200, which indicates that every hawker centre caters 200 people.

Ranking of classified field by using reclassify tool to rank the regions based on the hawker centre density which shows places ranked according to economic zones for food.
Based on proximity to cycling paths:

The areas are ranked as per its proximity of cycling paths and data is converted to raster data and the ranked.

Ranking MRT by reclassify:

The easy access to public transport is considered one of the major consideration while choosing a place to stay and have used proximity to MRT as a factor to rank areas. MRT location data is converted to raster data to rank areas based on its proximity to MRT stations.
Final Ranking:
The final rank for each zone in Singapore was calculated based on average of other three ranks (MRT, Hawker centre and Cycling path). Areas which are ranked good can be most favourable for staying. This final ranking can help to choose a place for staying based on individual priority.

Recommendations:
Using the final ranking, we can recommend to a new student coming to study at NUS a convenient place to stay, considering MRT, Cycling and Food places in the ranking.

As expected the better ranking zones are crowded near NUS itself and there are other places also being suggested by the ranking.

As a future scope of this story, we can add a configurable element which can replace Hawker centre layer with many other layer like Libraries, Parks, HDB rental prices, Bus stop layer to form an ideal tool for upcoming student to use it.
Appendix: (This analysis can be used to further improve the rankings of zone)

Libraries

Libraries and Nation Park after Spatial Join: