# Title: Evaluation of hospital catchment area algorithms as applied to Ho Chi Minh City

## Student: Weber Liu

SID: 450214701

Research Tutor: Dr. Justin Beardsley

Expert Advisor: Dr. Justin Beardsley

Word count

Cover statement (student): 169

Cover statement (research supervisor):

Abstract: (max 250 words)

Body: (max 3000 words)

### Cover statement – Research supervisor (Dr. Justin Beardsley)

### Cover statement – Weber Liu

The contents of the following report are solely of my own work. The development of the research question was done so in conjunction with my research tutor, Dr. Justin Beardsley. The initial stage of data de-identification was performed by my colleague, Mr. Joshua Chambers in order to comply with the ethical requirements for this project. Whilst the same dataset was used by Ms. Bebe D’Souza, Mr. Joshua Chambers and myself, there was no sharing of data whatsoever after the de-identification stage. The current project relates to the work of the other individuals of the research group insofar as the same data was used. Whilst the other individuals within this group have analysed a component of data, or have generated catchment area maps with a single algorithm, this project has generated generalised catchment area maps.

Originally, I had been given the 2015-2016 Hospital admissions data of Hospital for Tropical Diseases and Pham Ngoc Thach hospital from Dr. Justin Beardsley. Whilst the team had unanimously decided to use the Google Maps API for geocoding, I noticed that this decision was lacking an evidence base. After an initial literature review, I discovered that the usage of these geocoding APIs were well studied in English-speaking countries but severely lacking in Asian countries such as Vietnam. From here, I decided to take my MD projects back to the basics and document the complete procedure and attempt to build my own evidence base around why certain options were chosen over others. Through this, the results of the current project offer guidance regarding future choices for catchment area map generation in Vietnam.

Further, it was originally Dr. Beardsley’s idea to create a series of resources to provide researchers and clinicians in Vietnam describing the steps in producing catchment area maps as in the current project. As such, I have created open sourced Python scripts along with complete documentation and videos, currently hosted on my project website at <https://ouibaa.github.io/l/CatchmentAreas>.

The challenges I faced throughout this project was in the use of geocoding algorithms. As Google and HERE maps are commercial algorithms, the use of their algorithms costs money. This would pose a problem each time the process had to restart, for example, if my computer had shut down or if a bug in my code caused the software to crash. Eventually all analyses were performed without incurring cost. Further, as I do not read Vietnamese, I faced challenges in assessing geocoding accuracy, which I overcame by a slow visual comparison on a small subset (5%)

I would like to thank Dr. Justin Beardsley for supporting me throughout the project and providing greater insight regarding the concept of hospital catchment areas. I would also like to thank Dr. Justin Beardsley, Dr. Greg Fox. Professor Nicholas Manolios and A/Prof Andrew Bleasel for their feedback and engagement throughout the milestones for this project.

### Abstract

Background:

Hospitals rely on knowledge of their catchment areas for adequate planning of resource allocation and to help understand the local epidemiology of those utilising their services. True utilisation patterns should be consulted when deriving catchment areas for any hospital rather than simply assuming hospital choice is convenience based. Radius and distance-matrix based mapping approaches have previously been the mainstay approach due to convenience and lower computational requirements. Patient flow methods such as those utilising cumulative case ratios have since superseded such approaches and provide insight into who and how healthcare resources are used by.

Method:

We use hospital admissions data from Hospital for Tropical diseases (HTD) and Pham Ngoc Thach (PNT) across 2015-2016 to calculate their respective catchment areas. Geocoding APIs from Google, Bing and HERE are used to parse address data and their accuracies are assessed with Excel. The haversine formula is used to calculate admissision-hospital displacements to generate a radius-based catchment area and Google Maps Distance Matrix Algorithm is used to calculate travel time and duration to derive Distance Matrix-based catchment area maps. The 2015 WorldPop Vietnam dataset is used in the derivation of cumulative case ratio (CCR) based catchment areas.

Results:

Google maps was significantly more accurate in geocoding both HTD and PNT datasets compared to Bing and HERE maps. We were able to map and describe displacement, distance-matrix and CCR catchment areas and provide insight into the choice of case-rate for CCR-based catchment maps in Ho Chi Minh City.

### Introduction

**Hospital catchment areas**

The hospital catchment area, defined by Senn and Samson (1982) as the “group of persons who would attend the hospital or unit were they required to treatment” (1), provide hospitals with knowledge of their local epidemiology and allow planned allocation of resources to support their communities. Early catchment area served the simple purpose of indicating local population sizes (1), however with the increased use of Geographical Information Systems (GIS), we can now generate area-based maps and identify disparities in patient access to health services (2). Whilst initial models were simply based off simple displacement-from-hospital calculations, (3), improved computational power has enabled a more patient-centred approach or ‘patient flow’ models where segregated geographic areas can have their expected patient load contribution calculated for each hospital for higher resolution prediction of local health service utilisation (4).

**GIS and APIs**

The utility of Geographic Information Systems (GIS) in South-east Asian countries such as Vietnam is evidenced by its widespread applications, from surveillance and analysis of the Malarial outbreaks, to finding associations between neighbourhood quality and its resident’s health (5). As a public health tool, GIS has been used for the mapping and presentation of spatiotemporal data points, utilising population boundaries as potential sentinel nodes for predicting disease spread and oncoming endemics. GIS’s importance to hospitals lies in its potential to spatially map hospital incidence to elucidate a utilisation-based catchment area (6), allowing adequate preparations to be made for periods of high disease incidence.

In order to translate hospital data for calculation of catchment areas, address data must be first geocoded into a WGS84 (latitude and longitude) format. It is common for issues to arise in this process, with address data potentially incorrectly written or typed into admission forms, or investigators unable to locate an address due to unfamiliarity with local geography. As such, large database and software companies such as Google, Bing and TomTom have offered their mapping services for public use through their Application Programming Interfaces (APIs). These APIs are commonly used in epidemiological research however there is little evidence to guide our choice in API. We suspect that the platforms will differ in address resolution (e.g. recognising addresses to the unit level in an apartment complex) and the quality of their fuzzy-matching algorithm (i.e. capacity to deal with address spelling errors). Thus it is crucial to quantify geocoding API accuracy in order to guide future research.

Previously unfathomable computations such as prediction of driving route and travel times can additionally be calculated through access to Distance Matrix software via APIs, such as those used in Baidu maps (7) and OpenStreetMap (8). These allow more realistic calculation of catchment areas and provide a better understanding of convenience-based hospital utilisation.

In the current paper, we developed a catchment area map generation pipeline, as shown in Figure 1. We have assessed the accuracy of various geocoding APIs in relation to 2015-2016 admissions data from the Hospital for Tropical Diseases (HTD) and Pham Ngoc Thach hospital (PNT), Vietnam and describe metrics in this admissions data to guide future catchment area map generation for hospitals in Vietnam. We have also compared` three different catchment area algorithms based on displacement, distance matrix and a cumulative case ratio. The displacement algorithm included wards within a given radius from the hospital of interest and is the most simplistic. This is calculated using the Haversine formula (Appendix X). The distance matrix algorithm required the use of commercially available APIs and has generate catchment area maps based on both travel distance and travel times for admissions to their hospitals of interest (Appendix X). Finally, the cumulative case ratio algorithm is a statistical patient-flow based algorithm originally suggested by Zinszer et al (6) and reflects the wards which over- or under-utilise the hospital compared to the expected population rate (Appendix X).

Graphical user interface, application

Description automatically generated

**Figure 1: Catchment area map generation pipeline – PNT and HTD** hospital admissions data from 2015-2016 is provided to the researcher, who then undertakes geocoding and subsequently distance matrix calculations using commercially available APIs (e.g. Google, Bing, HERE, OSM). Python scripts provided in the Github repository for this project can then be used to calculate straight line (Haversine equation). With the use of WorldPop data, case rate and cumulative case ratio data can be calculated via QGIS, and can be plotted onto basemaps from GADM.   
PNT = Pham Ngoc Thach; HTD = Hospital for Tropical Diseases; OSM = OpenStreetMaps; GADM = Database of Global Administrative Areas; HCMC = Ho Chi Minh City;

### Aims

We wish to compare the geocoding accuracies between Google Maps, Bing Maps and HERE maps for the hospital admissions data from both HTD and PNT. We also aim to perform exploratory statistical analyses on the HTD and PNT admissions data to inform future catchment area derivation on Vietnamese hospital admissions datasets. Finally, we wish to generate catchment area maps based on displacement, distance matrix and cumulative case ratio approaches, and compare their predicted catchment areas.

### Methods

**Data sources**

The dataset has been retrospectively collected from admissions data from Hospital for tropical diseases (HTD), and Pham Ngoc Thach University of Medicine Hospital (PNT) over the course of 2015-2016. This data is de-identified, with information about patient date of birth, approximate address (in longitude and latitude), gender, admission and discharge dates, and ICD-10 coded discharge diagnoses. This data requires no ethics approval.

Local population data was obtained from WorldPop (<https://worldpop.com>) (9, 10) and analysed for the population level using QGIS v3.10. All generated data was in the WGS84 format (latitude and longitude) for cross-software compatibility.

**Geocoding**

The dataset was manipulated to fit the geocoding documentation with Excel (Build 12527.21330; Microsoft) Commercial licences were obtained for the Google Maps Platform (<https://developers.google.com/maps>) and HERE Developer platform (<https://developer.here.com/develop>) for use with their respective REST APIs. An educational licence was obtained for the Bing Maps platform (<https://www.microsoft.com/en-us/maps>) for use with their REST API. Python v3.9.0 (Python Software Foundation) was used to programmatically retrieve addresses from the dataset and query the Google, Bing and HERE maps REST APIs, and save the returned fuzzy-matched address and geocoded latitudes and longitudes into a separate spreadsheet. Excel was then used to combine geocoded data with the original address datasets.

**Assessment of geocoding accuracy**

Geocoding accuracy was determined by manual comparison between original address data and the geocoded output from Google maps, HERE maps, Bing maps and OpenStreetMaps Nominatim APIs. Geocoding completeness (match rate) was determined by calculating the proportion of returned errors or empty location data to the total number of addresses. APIs with a low match rate were excluded from further analyses.

A random sampling approach was used to retrieve 5% of the entries from HTD and PNT patient datasets. A blinded comparison was made visually in Excel to match geocoded addresses to original addresses at the street-number, street name or ward level where such information was provided in the original address.

**Map development**

Maps were developed in qGIS version 3.10 (QGIS development team), using base maps (geoPackage files) from Database of Global Administrative Areas (GADM) version 3.6. Data was first organised in Excel before being imported into qGIS as vector data. Separate maps are developed for Straight-line algorithm, Distance matrix algorithm and Cumulative case ratio algorithm. The respective latitude and longitude values for the HTD and PNT hospital locations as used in these algorithms was obtained through Google Maps.

The straight line distance algorithm was performed by use of the Haversine formula to calculate the distance between each admission and their respective hospitals (Appendix X). This process was automated through a Python script. The catchment areas were defined as wards with at least one admitting patient within 1km, 5km, 10km, 20km and 30km displacement from the hospital.

The distance matrix algorithm was applied by parsing WGS84 geolocated patient addresses through the Google Maps Distance matrix API, which returns an estimated travel time and driving road-distance between the admission source and their respective hospitals. Distance matrix estimates were retrieved on the 17th of July 2020, and estimated for Monday 20th of July 2020 at 12pm local time (+0700 GMT). Catchment areas were defined as wards with at least one admitting patient within 1km, 5km, 10km, 20km and 30km driving distance from the hospital. Additionally, analyses were performed with estimated driving times, and were defined as wards with at least one admitting patient with a driving time less than 15, 30, 60 and 120 minutes to the hospital on Monday the 20th of July, 2020, at 12pm Local time.

The cumulative case ratio is the ratio between observed and expected cases within the smallest administrative region, in our case the ward. The expected cases is derived by multiplying the local population by the case rate, which is calculated as the proportion of cases in a given population. Our method is modified from Zinszer et al., 2014 (6) as inclusion of a ward into a catchment area is defined as simply when the cumulative case ratio of a ward exceeds 1 - that is, more observed cases than expected based on the baseline population rate. Three cumulative case ratio maps were developed based on different baseline populations: All of Vietnam, Ho Chi Minh City only, and exclusively from wards with at least one admitting patient.

**Statistical analysis**

RStudio (Version 1.3.1093; RStudio) was used for statistical analyses. Proportional differences in geocoding accuracy and population capture for both HTD and PNT data was calculated using the two-proportion z-test.

**Resource development**

As part of this project, instructional documents and videos were produced to demonstrate the analysis of admissions data and production of catchment area maps for the given hospitals in Excel, QGIS, Python and R for use by local researchers and clinicians in Vietnam. Further, the Python and R scripts used for calculations and API queries have been made open source by the author. These files are available on the author’s website at <https://ouibaa.github.io/l/CatchmentArea>

**Ethics statement**

The ethics for this project were granted by hospital-specific ethics committees for HTD and PNT in Vietnam. As this was deemed LNR, no approval numbers were provided for such a study, and letters of permission were provided and signed by heads of these committees. Additionally, approval to conduct this project was also provided by Oxford University. However, as this study is not conducted on Australian datasets, no relevant HREC in Australia has approved this study.

### Results

**Geocoding accuracy**

**Figure 2: Geocoding accuracy for HTD (red) and PNT (black) admissions data using Bing, Google and HERE maps.**

Figure 2 compares the geocoding accuracies of Bing Maps, Google Maps and HERE Maps APIs for the HTD and PNT admissions datasets. Geocoding accuracy was greatest for both HTD and PNT using the google maps API (22.1% for HTD and 60.3% for PNT). Google maps more accurate than Bing in the geocoding of HTD data by 9.4% (95% CI 6.46% to 12.4%) ( (1df) = 38.107; p < 0.001) and in PNT data by 33.33% (95% CI 29.2% to 37.4%) ( (1df) = 225.41; p < 0.0001). Google maps was also significantly better than HERE in the geocoding of HTD data by 12.78% (95% CI 9.86% to 15.7%) ( (1df) = 72.111; p < 0.0001) and PNT data by 7% (95% CI 2.67% to 11.3 ( (1df) = 9.9847; p < 0.005). Bing geocoding accuracy was similar for HTD (25.4%) and PNT (27%). HERE maps performed better in geocoding accuracy for PNT data (53.3%) compared to HTD data (22.1%). Whilst Bing geocoding accuracy was significantly better than HERE for HTD data by 3.33% (95% CI 0.55% to 6.1%) ( (1df) = 5.5174; p < 0.05), HERE was significantly more accurate than Bing for PNT data by 26.3 % (95% CI 22.2% to 30.4%) ( (1df) = 143.92; p < 0.0001).

Match rate was high across all geocoding APIs for all data. HTD data could not match 0.27% of records for Bing, 0.22% of records for Google maps and 1.94% of records for HERE maps. Similarly, PNT data could not be matched for 0.1% of its data with Bing, 2.2% with Google maps and 3.5% with HERE. OpenStreetMaps Nominatim (<https://nominatim.org/>; OpenStreetMap foundation) had the lowest match rate of 69.8% and was thus not used for any further analyses.

**Exploratory statistics of dataset**

**HTD**

The median displacement of all 2015-2016 admissions to the HTD hospital was 10km (IQR 4.13km – 33.4km), with 90% of patients residing within a 127km radius. The median driving distance was further from the HTD hospital, at 14.5km (IQR 6.7km – 43.4km), with 90% of patients driving less than 156km to the hospital. The median driving time was 31 minutes (IQR 16.68 – 72.8 mins), with 90% of patients travelling less than 3 hours 28 minutes to reach the hospital.

**PNT**

Patients from PNT hospital are spread more widely across Vietnam compared to HTD. The median displacement of admissions to PNT hospital was 33.4km (IQR 6.7 – 11.3km) with 90% of patients within a 203km radius. The median driving distance was greater at 40.4km (IQR 8.7km – 144km, 90% within 260km). The median driving time was 71 minutes (IQR 23min – 185min), with 90% of patients travelling less than 5 hours 23 minutes to the hospital.

**Catchment area model characteristics**

**HTD**

Table 2 describes the characteristics of the admission data and catchment area model for HTD. Using a 10km displacement, the catchment area contains 229 wards and resides completely within HCMC (Figure 3A). Similarly, a 10km driving distance model contains 196 wards (Figure 3B), and a 30 minute driving time contains 227 wards (Figure 3C), all of which are contained within HCMC. The 1km driving distance model the least number of wards and the cumulative case ratio (VNM case rate) model has the most. Of the cumulative case rate models, the HCMC case rate model (Figure 3E) included the least number of wards in HCMC whilst the VNM case rate model (Figure 3F) included the most.

**PNT**

Table 3 similarly describes the characteristics of admission and catchment area models for PNT. As with HTD, the 10km displacement, 10km driving distance and 30 minute driving time models for each reside completely within HCMC, at 233 wards, 204 wards and 227 wards respectively. As with HTD, the 1km driving distance model is the smallest catchment area while the Cumulative Case Ratio (VNM Case rate) model is the largest. Of the cumulative case rate models, the HCMC case rate model (Figure 4E) is the smallest whilst the VNM case rate (Figure 4F) is the largest in terms of HCMC ward coverage. In general, All catchment areas derived from the PNT dataset include more wards than that of the HTD dataset with the exception of the 30km Displacement catchment area model.

**Table 2: HTD admissions data and models for Haversine, Network analysis (Google) and Cumulative case ratio**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **Admissions data** | | | **Model variables** | | | |
| **Admissions within bounds (% Total)** | **Number of source wards** | **Case Ratio (per 1000 people)** | **Catchment population (% Source)** | **Number of catchment wards (% Source)** | **Catchment population within HCMC (%HCMC)** | **Number of catchment wards in HCMC (% HCMC)** |
| **Haversine (Displacement model)** | |  |  |  |  |  |  |  |
|  | 1km | 2349 (6.05) | 16 | - | 308150 (0.33) | 16 (0.14) | 308150 (3.49) | 16 (4.97) |
|  | 5km | 13860 (35.72) | 150 | - | 2983618 (3.23) | 150 (1.34) | 2983618 (33.75) | 150 (46.58) |
|  | 10km | 21020 (54.17) | 229 | - | 6127260 (6.64) | 229 (2.05) | 6127260 (69.31) | 229 (71.12) |
|  | 20km | 26254 (67.66) | 326 | - | 9247040 (10.02) | 326 (2.92) | 8312251 (94.02) | 294 (91.3) |
|  | 30km | 28871 (74.40) | 437 | - | 11559372 (12.53) | 437 (3.91) | 8580339 (97.05) | 306 (95.03) |
|  | Total dataset | 38805 (100) | 3116 | - | - | - | - | - |
| **Network analysis model (Google maps)** | |  |  |  |  |  |  |  |
|  | 1km | 295 (0.76) | 2 | - | 36548 (0.034) | 2 (0.018) | 36548 (0.41) | 2 (0.62) |
|  | 5km | 7572 (19.54) | 79 | - | 1465505 (1.59) | 79 (0.71) | 1465505 (16.58) | 79 (24.53) |
|  | 10km | 17539 (45.26) | 196 | - | 4739012 (5.13) | 196 (1.76) | 4739012 (53.6) | 196 (60.87) |
|  | 20km | 24533 (63.31) | 291 | - | 8140512 (8.82) | 291 (2.61) | 8013724 (90.64) | 283 (87.89) |
|  | 30km | 27101 (69.93) | 353 | - | 9998775 (10.84) | 353 (3.16) | 8444896 (95.52) | 298 (92.55) |
|  | Total dataset | 38753 (100) | 3100 | - | - | - | - | - |
| **Network Analysis model travel duration (Google maps)** | |  |  |  |  |  |  |  |
|  | 15 min | 10792 (27.85) | 109 | - | 469021 (0.51) | 109 (0.98) | 469021 (5.31) | 109 (33.85) |
|  | 30 min | 10840 (27.97) | 227 | - | 1089239 (1.18) | 227 (2.03) | 1089239 (12.32) | 227 (70.5) |
|  | 1 hr | 27560 (71.12) | 394 | - | 2079681 (2.25) | 394 (3.53) | 1471845 (16.65) | 295 (91.61) |
|  | 2 hr | 32365 (83.52) | 912 | - | 7677554 (8.32) | 912 (8.17) | 1674672 (18.94) | 321 (99.69) |
|  | Total dataset | 38753 (100) | 3100 | - | - | - | - | - |
| **Cumulative case ratio model** | |  |  |  |  |  |  |  |
|  | VNM Case rate | 38805 (100) | 3116 | 0.421 | 17123860 (18.56) | 1173 (10.5) | 8291644 (93.79) | 302 (93.79) |
|  | HCMC Case rate | 24870 (100) | 322 | 2.813 | 3136985 (35.48) | 113 (35.1) | 3136985 (35.48) | 113 (35.09) |
|  | Patient Source Ward Case Rate | 38805 (100) | 3116 | 0.943 | 8714142 (9.44) | 514 (16.5) | 6368613 (72.04) | 244 (75.78) |

**Table 3: PNT admissions data and models for Haversine, Network analysis (Google) and Cumulative case ratio**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **Admissions data** | | | **Model factors** | | | |
| **Admissions within bounds (% Total)** | **Number of source wards** | **Case Ratio (per 1000 people)** | **Catchment population (% Source)** | **Number of catchment wards (% Source)** | **Catchment population within HCMC (%HCMC)** | **Number of catchment wards in HCMC (% HCMC)** |
| **Haversine (Displacement model)** | |  |  |  |  |  |  |  |
|  | 1km | 421 (2.22) | 22 | - | 251220 (0.27) | 22 (0.19) | 251220 (2.84) | 22 (6.83) |
|  | 5km | 4134 (21.75) | 153 | - | 3400105 (3.68) | 153 (1.37) | 3400105 (38.46) | 153 (47.52) |
|  | 10km | 6410 (33.73) | 233 | - | 6322588 (6.85) | 233 (2.09) | 6322588 (71.52) | 233 (72.36) |
|  | 20km | 8284 (43.59) | 326 | - | 9224004 (10.0) | 326 (2.92) | 8312251 (94.02) | 294 (91.3) |
|  | 30km | 9362 (49.26) | 436 | - | 11531983 (12.5) | 436 (3.91) | 8605139 (97.33) | 306 (95.03) |
|  | Total dataset | 19004 (100) | 3116 | - | - | - | - | - |
| **Network analysis model (Google maps)** | |  |  |  |  |  |  |  |
|  | 1km | 186 (0.98) | 10 | - | 124515 (0.13) | 10 (0.09) | 124515 (1.41) | 10 (3.11) |
|  | 5km | 1981 (10.45) | 110 | - | 2231694 (2.42) | 110 (0.99) | 2231694 (25.24) | 110 (34.16) |
|  | 10km | 5485 (28.93) | 204 | - | 5178525 (5.61) | 204 (1.83) | 5178525 (58.57) | 204 (63.35) |
|  | 20km | 7880 (41.56) | 299 | - | 8573587 (9.29) | 299 (2.68) | 8159400 (92.29) | 287 (89.13) |
|  | 30km | 8724 (46.01) | 363 | - | 10117282 (10.96) | 363 (3.25) | 8484569 (95.97) | 300 (93.17) |
|  | Total dataset | 18961 (100) | 3096 | - | - | - | - | - |
| **Network Analysis model travel duration (Google maps)** | |  |  |  |  |  |  |  |
|  | 15 min | 3093 (15.31) | 117 | - | 2591333 (2.81) | 117 (1.05) | 2591333 (29.31) | 117 (36.34) |
|  | 30 min | 6092 (32.13) | 227 | - | 6016179 (6.52) | 227 (2.03) | 6016179 (68.05) | 227 (70.5) |
|  | 1 hr | 9071 (47.84) | 384 | - | 10503416 (11.38) | 384 (3.44) | 8459549 (95.69) | 298 (92.55) |
|  | 2 hr | 12432 (65.57) | 915 | - | 17334210 (18.78) | 915 (8.20) | 8831174 (99.89) | 321 (99.69) |
|  | Total dataset | 18961 (100) | 3096 | - | - | - | - | - |
| **Cumulative case ratio model** | |  |  |  |  |  |  |  |
|  | VNM Case rate | 19004 (100) | 3116 | 0.20594 | 27835370 (30.16) | 2104 (18.84) | 8744070 (98.91) | 318 (98.76) |
|  | HCMC Case rate | 8209 (100) | 322 | 0.92853 | 3344134 (37.83) | 168 (0.52) | 3344134 (37.83) | 168 (52.17) |
|  | Patient Source Ward Case Rate | 19004 (100) | 322 | 0.45767 | 14369510 (15.6) | 1052 (9.42) | 7198219 (81.42) | 227 (70.5) |

|  |  |  |
| --- | --- | --- |
| **A** | **B** | **C** |
| **Map  Description automatically generated** | **Map  Description automatically generated** | **Map  Description automatically generated** |
| **D** | **E** | **F** |
| **Map  Description automatically generated** | **Map  Description automatically generated** | **Map  Description automatically generated** |

**Figure 3: HTD catchment area maps derived by (A) straight line (haversine) distance, (B) Google Maps Distance matrix for driving distance, (C) Distance matrix for driving time, (D) Cumulative case ratio with the case rate of wards with at least 1 admission, (E) Cumulative case ratio with the HCMC case rate and (F) Cumulative case ratio with the VNM case rate**

|  |  |  |
| --- | --- | --- |
| **A** | **B** | **C** |
|  |  | **Map  Description automatically generated** |
| **D** | **E** | **F** |
|  |  |  |

**Figure 4: PNT catchment area maps derived by (A) straight line (haversine) distance, (B) Google Maps Distance matrix for driving distance, (C) Distance matrix for driving time, (D) Cumulative case ratio with the case rate of wards with at least 1 admission, (E)**

**Cumulative case ratio with the HCMC case rate and (F) Cumulative case ratio with the VNM case rate.**

### Discussion

**Geocoding accuracy**

Geocoding accuracy was problematic for both PNT and HTD hospital datasets and for all three API providers used in this study. Most issues arose from inconsistencies in how addresses were recorded at the hospital side, however this issue was likely also compounded by inadequate handling of non-English address data by Google, Bing and HERE, which primarily service English-speaking countries.

The main issue observed in assessment of accuracy arose from incorrectly formatted data. Many included addresses did not include street numbers, or only included the name of a road which may have stretched across multiple wards. Shorthand notations such as ‘TT’ for ‘Thi Tran’ and ‘QL’ for ‘Quoc Lo’ appeared inconsistently in both patient address data and in geocoding output. Another common issue involved the geocoding APIs confusing apartment or unit numbers with street numbers and returning a very similar address in a different ward or location to the provided details.

The reported geocoding error rate in this study likely overestimates the true error rate of the geocoding services as the authors are unable to read Vietnamese, and so a systematic method of visual comparison was employed for the assessment stage. We acknowledge this approach could incorrectly reject correctly geocoded addresses presented in format dissimilar to that provided in the admissions dataset.

Google maps consistently offered the best geocoding accuracy whilst HERE maps was better for PNT but worse for HTD data. This may be due to HERE maps’ inability to support high resolution data (e.g. it cannot identify to the unit-number level within an apartment complex), which may explain why HERE maps performs better for PNT, where more patients reside outside the densely packed HCMC city area. However, subgroup analyses were not performed to assess this claim and should be considered in further studies. In comparison, Bing maps has poorer resolution compared to the other two services. OpenStreetMaps Nominatim was not included for any analysis after it was assessed to have a high error rate, likely due to an inferior fuzzy matching algorithm compared to Google, HERE and Bing maps. This may be overcome by using alternate providers of OpenStreetMaps (e.g. TomTom). Inaccurate geocoding issues may also be circumvented from the hospital side by implementation of a geocoding API at the point of patient admission to ensure only recognisable addresses are included in the admission records.

**Admissions data**

The spread and diversity of patients attending PNT and HTD hospitals as demonstrated within our exploration of dataset should inform future map production and choice of displacement and distance matrix boundaries. Interestingly, PNT hospital had a larger spread of patients across Vietnam compared to HTD despite less admissions. This is possibly due to factors such as increased specialisation or patient preference but may also be a result of highly inaccurate geocoding.

**Catchment area models**

Traditional displacement-based and distance matrix models are calculated without consideration for the source of patient, thus resulting in concern that such models are overgeneralised or simplified. Such is the reason that statistical and patient-flow methods such as the Cumulative Case Ratio methods presented in this paper and in Zinszer et al 2013 (6). In all models presented here, the catchment areas consist of only wards from where there is at least one admitting patient. Whilst this overcomes the issue of overgeneralisation, these maps are more resource intensive to produce and are non-contiguous. These maps may therefore also vary between years as different populations of patients attend the hospital or as road and highway infrastructure develops. Moreover, this is the first paper we have identified which has utilised a distance matrix API to predict real-time traffic data in the generation of a distance matrix driving-time and driving-distance catchment area map.

Cumulative care ratio models are clearly non-contiguous and three separate maps based on different population case rates were developed to assess the best model for differentiation of wards with major or minor contributions to the hospital. Whilst the VNM case rate models (Figure 3F and 4F) covered the most of HCMC (>90% of wards included in both models), these effectively suggest that all of HCMC should be considered as the catchment area. Rather, we believe the HCMC case rate models (Figure 3E and 4E) provide the most detail regarding high and low-contributing wards, and should be used as the preferred method of deriving cumulative case ratio catchment area models.

We acknowledge that all models presented in this study likely overestimates the true catchment area as the criteria for ward inclusion is for the smallest value to fall within bounds of each strata. Additionally, the use of the Haversine formula for displacement calculation assumes Earth is spherical and can be improved with the spheroid assumption in Vincenty’s formula. The effect of such differences may be of interest in future investigations.

### Conclusion

In the current study we generated three maps based on patient displacement, distance matrix and cumulative case ratio for two separate hospitals in Ho Chi Minh City, Vietnam. We determined that Google Maps performed better than Bing and HERE maps in geocoding Vietnamese address data, and generated the first known catchment area map based on predicted real-time driving times using the Google Maps API. Finally we assessed three separate version of the cumulative case ratio catchment area algorithm based on different base population rates and determined the use of an isolated Ho Chi Minh City case rate provides the greatest discrimination for ward inclusion in our datasets.

### References

1. Senn SJ, Samson WB. Estimating Hospital Catchment Populations. The Statistician. 1982;31(1).

2. McLafferty SL. GIS and health care. Annu Rev Public Health. 2003;24:25-42.

3. Phibbs CS, Robinson JC. A variable-radius measure of local hospital market structure. (0017-9124 (Print)).

4. Gilmour SJ. Identification of hospital catchment areas using clustering: an example from the NHS. Health Serv Res. 2010;45(2):497-513.

5. Bui TQ, Pham HM. Web-based GIS for spatial pattern detection: application to malaria incidence in Vietnam. (2193-1801 (Print)).

6. Zinszer K, Charland K, Kigozi R, Dorsey G, Kamya MR, Buckeridge DL. Determining health-care facility catchment areas in Uganda using data on malaria-related visits. Bull World Health Organ. 2014;92(3):178-86.

7. Tao Z, Yao Z, Kong H, Duan F, Li G. Spatial accessibility to healthcare services in Shenzhen, China: improving the multi-modal two-step floating catchment area method by estimating travel time via online map APIs. BMC Health Serv Res. 2018;18(1):345.

8. Lindholm D, James S, Lagerqvist B, Hlatky MA, Varenhorst C. New Method for Assessing the Effect of Driving Distance to Hospital Care: Using OpenStreetMap Routing in Cardiovascular Research. Circ Cardiovasc Qual Outcomes. 2017;10(9).

9. WorldPop. Viet Nam 100m Population: School of Geography and Environmental Science, University of Southampton; 2013 [Alpha version 2010, 5 and estimates of numbers of people per pixel (ppp) and people per hectare (pph), with national totals adjusted to match UN population division estimates (<http://esa.un.org/wpp/>) and remaining unadjusted.]. Available from: <www.worldpop.org>.

10. Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ. High resolution population distribution maps for Southeast Asia in 2010 and 2015. PLoS One. 2013;8(2):e55882.

### Acknowledgements

We would like to acknowledge Associate Professor Andrew Bleasel, Dr. Greg Fox and Professor Nicholas Manolios for their contributions and suggestions regarding the direction of this project. We would like to also acknowledge Microsoft Bing for providing a research licence allowing for unlimited use of their mapping software.

### Appendices

**Appendix 1: Python scripts for automated geocoding**

The following scripts are open-source and available at <https://github.com/weberliuMD/Hospital-Catchment-Area-Tools-H-CAT>

## Appendix X – Haversine formula for displacement calculation

The haversine formula is used to calculate the straight-line distance between two points on Earth, given knowledge of their latitude and longitudes, where Earth is assumed to be spherical.

The straight-line distances were calculated between the hospital locations ([10.753167,106.678477] and [10.756397, 106.665468] for HTD and PNT respectively) and geocoded patient addresses. The calculation of straight line distance was automated through a Python script.

## Appendix X – Distance Matrix

## Appendix X – Cumulative case rate