

Master Thesis

submitted within the UNIGIS MSc programme
Interfaculty Department of Geoinformatics - Z_GIS
University of Salzburg

**Google *Street View*, SANET kernel density estimation, and
linguistic diversity of the Malaysian linguistic landscape**

by

B.A. Ryan Scamehorn

U103606

ryan.scamehorn@gmail.com

A thesis submitted in partial fulfilment of the requirements of
the degree of
Master of Science (Geographical Information Science & Systems) – MSc (GISc)

Advisor:

Dr. Gudrun Wallentin

A B S T R A C T

The main objective of this thesis is to use Google *Street View* to collect linguistic landscape samples then use network kernel density estimation to visualize the network location patterns of multilingual relationships in selected regions and scales of Malaysia's linguistic landscape and then compare results using co-location patterns from additional KDE measures in analogous study regions and scales; secondary objective is to discuss the theoretical importance of *K*-functions in analogous regions. Linguistic landscape is well-suited to the study of the geographic distribution of multilingualism because "one can use the linguistic landscape as an indicator of the power and status relationships that exist between the various language groups present within a given administrative or geographical region." (Bourhis, Landry, 1997) Throughout peninsular Malaysia, the linguistic landscape offers a mine of data of keen interest to linguistic geographers running in vast tapestries of street-level photographic coverage. Malay, English, Tamil, Chinese, and Arab-Jawi scripts are clearly present in mono-, bi-, and polylingual configurations, now freely available for high-volume data collection through supervised geo-tagging. By employing Google *Street View*, data collection of the along-network linguistic tokens is made easier and cheaper, producing big datasets well-suited to the analytical powers of a GIS. Malaysia presents an intriguing combination of 'inclusive' and 'exclusive' scripts at a high national linguistic diversity measure of .758 (UNESCO, 2009). Because linguistic landscape tokens are an along-network function of the transportation routes on which they are observed, so linguistic landscape datasets should be analyzed in a non-Euclidean space. Of all the point pattern analysis techniques, kernel density estimation is used to illustrate the changes in *lingua franca* preference at local, district, and state scales by using a limited non-intersected network. The network kernel density estimation method analysis toolkit is made possible by the SANET network toolkit and research based on the work of Atsuyuki Okabe et al. Results for this study indicate that both large and small kernel density estimation bandwidths are optimal for visualizing *lingua franca Preference* and for visualizing these trends through the use of co-location patterns, done by double-projecting the kernel density measures along the same network link. Additionally, the linguistic diversity index calculated for Google *Street View*-collected linguistic landscape data was 0.740, a result close to the UNESCO World Report calculation for Malaysia at 0.758.

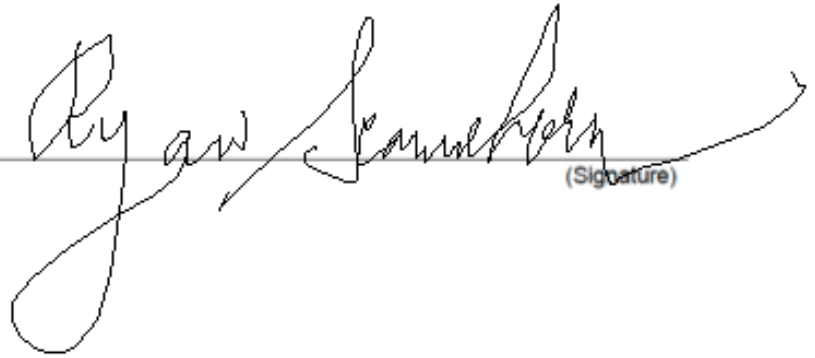
Science Pledge

By my signature below, I certify that my thesis is entirely the result of my own work. I have cited all sources I have used in my thesis and I have always indicated their origin.

Michigan, USA

15 April 2017

(Place, Date)

A handwritten signature in black ink, written in a cursive style, that reads "Ryan Sankharya". The signature is written over a horizontal line that spans the width of the signature area.

(Signature)

Introduction.....	8
1.1 Background.....	8
1.2 Objective.....	8
1.3 Approach.....	8
1.4 Expectations.....	10
1.5 Audience.....	10
1.6 Structure.....	10
Literature Review.....	12
2.1 GIS and Linguistics.....	12
2.2 GIS, Multilingualism, and linguistic diversity.....	14
2.3 GIS and the Linguistic landscape.....	15
Methodology.....	17
3.1 Overview.....	17
3.2 Multilingualism indicator design.....	20
3.3 Data collection method.....	24
3.4 Kernel density estimation and <i>K</i> -function.....	28
3.5 SANET Network kernel density estimation.....	31
3.6 GIS Software and Resources.....	35
Application.....	36
4.1 Sample Application for a Linguistic Landscape Analysis.....	36
4.1.1. Google <i>Street View</i> : Geo-tag Pointset and Export KML.....	37
4.1.2 Point Import, Point Decomposition.....	38
4.1.3 OSM Map Export.....	39
4.1.4 Network Building.....	40
4.1.5 Map Matching.....	40
4.1.6 SANET: Kernel Density Estimation.....	41
4.1.7 Base Heights Adjustment and Extrusion.....	42
4.1.8 Overlay and Transparency.....	42
Results and Analysis.....	43

5.1 Sample application results.....	43
5.1.1 Total tokens: Decomposition of the linguistic landscape.....	44
5.1.2 Total instances.....	45
5.1.3 Monolingual composition.....	46
5.1.4 Bilingual Co-occurrence.....	47
5.1.5 Affiliation.....	48
5.1.6 Greenberg’s Linguistic Diversity Index.....	49
5.2 Trend Map: Kuala Pilah.....	50
5.3 Trend Map: Tampin.....	51
5.4 Trend Map: Melaka.....	52
5.5 Trend Maps: Composite.....	53
Conclusions	56
6.1 Summary	56
6.2 Constraints.....	56
6.3 Further research.....	58
Bibliography	59

LIST OF FIGURES

THESIS STRUCTURE.....	11
AGGLOMERATION OF MULTILINGUAL RELATIONSHIPS.....	20
DENGUE FEVER PSA.....	22
LINGUISTIC LANDSCAPE INDICATOR DESIGN.....	23
LINGUISTIC LANDSCAPE SAMPLES A-M.....	24-7
PLANAR KERNEL DENSITY ESTIMATION FORMULA.....	27
LINEAR KERNEL DENSITY ESTIMATION FORMULA.....	31
‘MAINSTREETING’ WEIGHTING THRESHOLD.....	32
NETWORK <i>K</i> -FUNCTION MODEL.....	33
PLANAR vs NETWORK <i>K</i> -FUNCTION.....	34
APPLICATION WORKFLOW.....	36
DATA COLLECTION IN GOOGLE <i>STREET VIEW</i>	37
LINGUISTIC LANDSCAPE KML INSTANCE.....	37-8
GOOGLE EARTH KML EXPORT.....	38
QUERY BUILDER SQL EXPRESSION.....	39
<i>OPEN STREET MAP</i> EXPORT.....	40
MAP MATCHING ERROR.....	41
KERNEL DENSITY ESTIMATION.....	43
CO-OCCURRENCE CO-LOCATION VISUALIZATION AT 100-M BANDWIDTH.....	43
DECOMPOSITION OF LINGUISTIC LANDSCAPE GRAPH.....	43
TOTAL INSTANCES.....	45
MONOLINGUAL COMPOSITION.....	46
BILINGUAL CO-OCCURRENCE.....	47
AFFILIATION.....	48
LINGUISTIC DIVERSITY INDEX.....	49
MAP: LINGUA FRANCA PREFERENCE–KUALA PILAH.....	50
MAP: LINGUA FRANCA PREFERENCE–TAMPIN.....	51
MAP: LINGUA FRANCA PREFERENCE–MELAKA.....	52
MAP: TOTAL LINGUISTIC LANDSCAPE COUNT (BI::BC)–MELAKA.....	53
MAP: COMPOSITE LINGUA FRANCA PREFERENCE.....	54

MAP: APPLICATION WINDOW: LINGUA FRANCA PREFERENCE.....	55
EXTRAPOLATION OF FULL SCALE LINGUISTIC LANDSCAPE SURVEY.....	56
GOOGLE OCR RESULTS ON LINGUISTIC LANDSCAPE TOKENS.....	57-8

I N T R O D U C T I O N

1.1 Background

While the geographic representation of linguistic data has long been a practice of linguists, the utilization of GIS by linguists is relatively new. “[Linguistics] has made little use of the powers of GIS...in spatial variation of language, correlated physical and social variables...the analytic and data processing capabilities are seldom discussed.” (Hoch and Hayes, 2010:23) The *linguistic landscape* presents an especially promising data source for examining the geographic distribution of languages and multilingualism due to its semi-permanence in both time and place. Linguistic landscape “serves to delineate the territorial limits of the language group it harbors relative to other linguistic communities inhabiting adjoining territories.” (Bourhis and Landry, 1997:25) It is the “written language of shop windows, commercial signs, official notices” (Gorter, 2006:1) that is ideally suited for high quantity data collection. In the study of the geographic distribution of multilingualism, “one can use the linguistic landscape as an indicator of the power and status relationships that exist between the various language groups present within a given administrative or geographical region.” (Bourhis, Landry, 1997) Throughout peninsular Malaysia, the linguistic landscape offers vast tapestries of street-level photographic coverage of Malay, English, Tamil, Chinese scripts, and Arab-Jawi scripts, clearly present in mono-, bi-, and polylingual configurations, now freely available for high-volume data collection through supervised geo-tagging. Malaysia presents an intriguing combination of ‘inclusive’ and ‘exclusive’ scripts at a high national linguistic diversity measure of .758 (UNESCO, 2009). By employing Google *Street View*, data collection of the along-network linguistic tokens is made easier and cheaper, capable of producing big datasets well-suited to the analytical powers of a GIS. As a function of the transportation routes on which they are observed, so linguistic landscape datasets should be analyzed in a non-Euclidean space. Kernel density estimation measures are used because they are useful at finding local and global trends. While especially useful at finding hotspots, directionality, and spatial mean centers, it the co-location pattern of double-projected KDE measures that are used in this study. Additionally, the theoretical use of network *K*-functions is discussed. They perform across a spectrum of scales allowing the detection of clustering and dispersion of selected linguistic configurations at the local, district, state, and national scale. Network kernel density estimation methods and network *K*-function methods are made possible by the SANET network toolkit and research based on the work of Atsuyuki Okabe.

1.1 Objective

The objective of this thesis is to include a Sample Application procedural use of *Street View* for linguistic landscape data collection and to use SANET kernel density estimations to analyze the co-location patterns of multilingual relationships in Malaysia’s linguistic landscape over a selection of regions and scales, and to then compare results to aggregate measures of linguistic diversity in analogous study regions and scales; a secondary objective is to discuss the theoretical importance of *K*-functions in analogous regions.

1.3 Approach

The motivation for this research is two-strand. First, though the use of *Street View* has been suggested by Barni and Bagna (2009) and later by Inoue (2012), there is an apparent disparity in the literature in which *Street View* is implemented as a tool for linguistic landscape data collection practices intended for spatial analysis by GIS. The second strand is to address the disparity in the GIS literature where SANET network kernel density estimation measures are employed to analyze linguistic landscape datasets.

Because the objective of this thesis is to analyze the linguistic landscape exclusively to generate a dataset for the study of multilingualism and linguistic diversity, Backhaus' definition of 'instance' is used to define the linguistic landscape instance. According to Backhaus, "any piece of text within a spatially definable frame" constitutes a single instance, from "small handwritten stickers to huge commercial billboards." (Gorter, 2006). The method used for calculating linguistic diversity is adapted from Greenberg's *Split Personality Method*, in which it is assumed that some speakers will be bi- and multilingual. However, in the case of the linguistic landscape, the 'speaker' is replaced by 'linguistic token'. This substitution assumes the assertion by Gorter that "the linguistic landscape is not a true reflection of the diversity of...in this perspective they refer to the linguistic landscape as the symbolic construction of the public space." (Gorter, 2012:195) Liao and Petzold (2010) and Graham and Zook (2013) also utilized a non-speaker formula for calculations of linguistic diversity. A set of custom metrics have been adapted for calculating the aggregate measures of *monolingualism, co-occurrence, and Preference*.

The decision to select Malaysia as the study region is manifold. It has a high linguistic diversity measure of .758 (UNESCO, 2009) Malay, English, Tamil, Chinese, and Arab-Jawi scripts are clearly present in mono-, bi-, and polylingual configurations throughout peninsular Malaysia. The linguistic landscape dataset used in this study is sampled exclusively from *Street View* street-level photographic coverage areas from the peninsular Malaysian states of Melaka, Negeri Sembilan, Pahang, and Kelantan. The clarity and extent of *Street View* coverage played a key role in the selection process.

Though Cenoz and Gorter (2008) propose up to sixteen criteria for evaluating the linguistic landscape, this study focuses on the role of multilingualism and linguistic diversity. Thus, only the languages(scripts) are sampled; no other criteria are recorded in the dataset or used in the spatial analysis. It should be noted that a number of Aslian languages are present in both peninsular Malaysia and Malaysian Borneo, though there is a negligible presence of Aslian in the linguistic landscape. Due in part to the remoteness as well as a lack of *Street View* coverage in Aslian speaking areas, the Orang Asli languages are not included as part of the linguistic agglomeration.

Though it should be noted that significant Orang Asli groups populate one study area, particularly N22 Bebar, Pahang, at 28% (UNDIInfo, 2013), it is beyond the scope of this thesis to analyze ethnographic or demographic data in correlation with the findings of the linguistic landscape, either through GIS functionality or through SANET tools.

QGIS was chosen as the GIS software for this thesis because the cross-platform open source software offers is a powerful free option that, together with *Street View* and SANET, offer a cost-free package of tools democratically available for the study of the linguistic landscape. The SANET toolkit as well as the theoretical research used in this study is the work of Atsuyuki Okabe. The SANET toolkit is available for use for academic and educational

use only. (SANET, 2016)

The spatial reference system in all datasets and shapefiles used in this study are EPSG:4326 (WGS 84), the SRS used by Google *Earth*.

1.4 Results

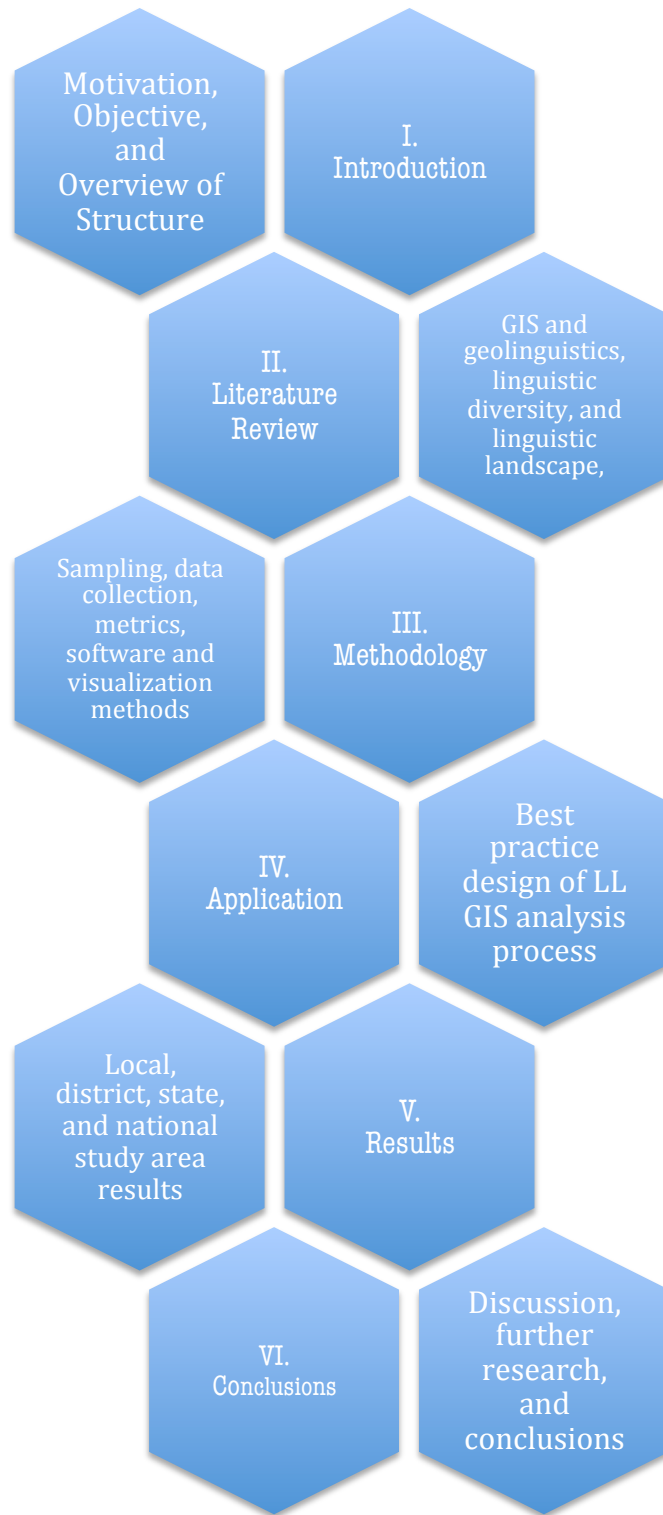
Due to an apparent lack of sampling methods with *Street View* or Sample Application outlining its use in linguistic landscape data collection, it is the intention of the author to describe the sampling process in detail so as it can stand alone as instructional guide to the process. It is the intention that this document should serve to guide the process from site selection, geocoding tags, through the export of KML and production of a trend map.

To reiterate the objective of this thesis: To “analyze the co-location patterns of multilingual relationships in Malaysia’s linguistic landscape over a selection of regions and scales, and to then compare results to aggregate and custom measures of linguistic diversity in analogous study regions and scales.” These patterns include: measuring levels of *monolingualism*; identifying *lingua franca* co-location patterns for various configurations of Malay (#BM) and English (#BI). These include intersecting and non-intersecting, monolingual and non-monolingual. These patterns are then tested in six small towns (pops. 18,000-58,000) one urban area (pop. 484,000), two state-wide scales, along a single network corridor. Results are generated into network trend maps using stacked ling graphs which visualize results as two- and three-color stacked line graphs. KDE measures are visualized as z-lines with an OSM base map.

1.5 Audience

Though the main objective of this work to use network kernel density estimation on linguistic landscape data for the GIS community, it is the intention of the author to address current issues in the linguistic literature, especially in issues related to the linguistic landscape and mapping methods, as well as indicators for linguistic diversity. Such a Sample Application may be of benefit to an audience of linguists, language policy makers, educators, or other academic or educational users.

1.6 Structure



THESIS STRUCTURE

L I T E R A T U R E R E V I E W

2.1 GIS and Linguistics

It is hard to imagine a time before the internet and personal computing in which the tasks of geolinguists could have been undertaken without descent into clerical nightmare. Imagine: the hand-wringing tedium of geo-referencing thousands of points from hand-written records taken from interviews and paper field notes without a GIS; or, compiling first-, second- and third-hand accounts across languages into atlases without employing the functionality of a geodatabase; or, tediously collecting data from paper mail responses without the convenience of email or online surveys; or, spending long hours in archives to make a single data request instead of a near-instant response from a government portal. One can readily infer that GIS and GIS spatial analysis must have only become widely available to linguists with the availability and accessibility to Big Data datasets and their data products, (TIGER shapefiles), the democratization of the GIS software (especially open source QGIS), and ubiquitous high-speed internet connections. Hoch and Hayes offer a blunt statement about the adoption of GIS by linguists:

In a review of the literature on geolinguistics, we found few studies either employing GIS or discussing methodology for doing so. Although researchers have developed GIS methods for spatial language data analysis, they do not often cite the history and progress of this development in the geolinguistics literature. (Hayes, Hoch, 2010) quoted from (Briscoe, 2009)

It was Van der Merwe's (1992) study of Cape Town that first used the term 'geolinguistics' to describe the field as an 'interdisciplinary' one, and he pointed out a lack of literature describing methods and history. Indeed, what may be the case with many interdisciplinary fields is that there seem to be a redundancy in terms. In a study by Williams (1996) from Briscoe (2009),

" 'Geolinguistics', 'language geography', 'geographical linguistics' or 'linguistic geography' used across the scholarly landscape is symptomatic for the diffused outline and origin of this field. I have come across linguistic geography being referred to as "an independent discipline in linguistics" (Fukushima and Heap 2008: 138), as well as Geolinguistics being called "an evolving branch of human geography" (Williams 1996: 63)

Whether geolinguistics is a branch of linguistics that employs geographic analysis or a linguistic-themed geography, researchers agree it is a nascent interdisciplinary field that is poorly researched and documented (Van der Merwe, 1992).

Van der Merwe's (1992) study of Cape Town set the tone for geolinguistics and their use of GIS. The study mentions GIS and how it is a tool that had a potential to serve the emerging interdisciplinary field. The analysis methods undertaken in this study determined the 'total surface area' of a spoken language within Cape Town, the 'Segregation Index' of selected languages, and an historical analysis that determined the shift of a language's 'Mean Centre of Gravity' using historical data and charts the change over the course over a single decade. The analyses utilized language data from a decennial census, focused on a single city, and produced choropleth maps.

Lee and Kretzschmar (1996) employed the use of the LASMAS Linguistic Atlas dataset to cartographically represent dialect data. They utilized point pattern analysis,

overlay functions, and spatial autocorrelation, visualizing results in isoglosses, as well as both Thiessen polygons and Delauney triangles. In addition, spatial statistics were employed. They also employed the use of spatial statistics, for which they posited “such as join count statistics can be used to validate the significance of spatial clustering, dispersion, or randomness” (Lee, Kretzschmar, 1996:541-59)

Veselinova and Booze followed Van der Merwe’s method of using census data, focusing on a single city, and producing choropleth maps, this time in Detroit. In their 2008 study, GIS was used to map Greenberg’s Linguistic Diversity, language density, spatial distribution, and segregation. Overlay analysis was employed to find patterns between language groups and property costs/rent, as well as correlate schools with foreign language instruction and income. (Veselinova, Booze, 2009:14).

Leubbering, Kolviras, and Prisley (2013) followed the Veselinova study featuring more in-depth spatial analysis in the greater Washington, D.C. area. Again a study of linguistic diversity, Greenberg’s Linguistic Diversity Index was rendered as a 3-D choropleth, as well as conversion of the vector census tracts to raster cells, in which the raster surfaces were smoothed by nearest neighbor analysis. (Leubbering, Kolviras, and Prisley, 2013:588) 3-D visualization may have been a poor choice to represent the linguistic diversity indicator due to the effect of obfuscation caused by some tall polygons.

Point pattern analysis was used with linguistic datasets by Lee and Kretzschmar in a 1993 study of American dialects. LAMSAS linguistic atlas pointsets were generated into Delauney triangles and Thiessen polygons. Their results were visualized with choropleth maps, from which join count statistics allowed for spatial autocorrelation. LAMSAS survey points were taken at data locations across Euclidean space at regular intervals, creating a near grid-like distribution of points that easily render into Thiessen polygons. (Lee, Kretzschmar, 1993) Point-to-polygon methods are not appropriate for linguistic landscape datasets due to the absence of tokens outside of transportation networks; LL tokens are almost exclusively a network edge phenomena. Google Street View these limitations. An additional problem lies in visualization of edge datasets as planar occurrences. As data is collected from the edges and nodes in a network, the question then becomes whether to split the edge down the middle of the street in to visualize polygons concurrently with administrative or cadastral units, such as a city block. This would require splitting linguistically distinct edges, dividing linguistically similar street from their opposite side and aggregating them with dissimilar ones. Conversely, one may keep the similar edges intact and visualize the street segment as a *dual graph*, though it would lose the native shape of administrative polygons, and thus some degree of GIS computational functionality. (Boeing, 2016)

After a thorough search, I was unable to find a single study that mentions spatial analysis of the linguistic landscape. In all the literature, the author was unable to find any mention of using linguistic data as a function of a network nor of the use of GIS network K -functions on linguistic datasets. The multilingual LL of Malaysia presents a challenging task for analysis. This is true not simply because of the plethora of configurations rendered from five written scripts, but the task of selecting from the wide array of available tools is a complex task in itself.

2.2 GIS, Multilingualism, and Linguistic Diversity

There are several ways to measure linguistic diversity. One measures the number of speakers in a given geographic space; another calculates the number of instances a language occurs in physical space and in cyberspace. A third is the linguistic landscape.

A well-known indicator for linguistic diversity comes from Greenberg's 1956 work, in which a number of methods for tabulating linguistic diversity are outlined. GIS-based studies by Leubbering et al (2013), and Veselinov et al (2008), employed Greenberg's first models, known as a *Monolingual Non-weighted Method*. Like the name indicates, the calculation counts the speaker only once, and its use implies a speaker only speaks a single language. However, Greenberg's includes additional metrics for polylingualism including the *Split Personality Method* (Greenberg, 1956:111) . It counts speakers more than once, a calculation indispensable if the true measure of linguistic diversity is to incorporate polylingual speakers in a measure of linguistic diversity. However, in order to make such a calculation, additional language data would be required of each participant. Because the census is not going to include additional language questions to accommodate such a need, census datasets are not well suited for linguistic diversity calculations where there are many polylingual speakers. The Malaysian census ceased to include language questions after 1980, so Greenberg's first method is no longer a viable option for use in Malaysia, regardless of whether it were to include questions to accommodate the high degree of polylingualism. (Lindsey, 2003)

New methods of calculating linguistic diversity employ the method of collecting aggregate language data from search engines. Such a method was pioneered by Funredes and Union Latine in the 1990s at a time when there was a greater amount of English language content on the internet than other languages. (Pimenta, Prado, Blanco, 2009) Since this study, greater degrees of accuracy and greater amounts of non-English content have added a greater degree of measureable linguistic diversity to the internet. However, early metrics simply gave a measurement of total internet linguistic diversity and did not geographically decompose the linguistic assessment.

Liao and Petzold (2010) utilized the method of mapping the aggregate content on Wikipedia. Though the organizational structure of Wikipedia allows for a measure of polylingualism in their data collection of linguistic diversity, Wikipedia content is not developed enough across all regions, which frustrates its use as a spatial dataset.

Later features by Google included geospatial search, and hence a spatial-based searches of aggregate internet content became possible with this tool. In a study by Graham and Zook (2013), this method was employed to map linguistic diversity at a number of resolutions in a number of linguistically conflicted areas: Hebrew::Arabic in Jerusalem, Spanish::Catalan in Spain, English::French in Quebec, and Dutch::French in Belgium. Their method of collecting aggregate search engine data visually conveyed spatial data in much the same way as a raster. A linguistic diversity measure was performed at regular spatial intervals. Each search point functions as a pixel in a rasterization and was visualized as a univariate map with a ramp between two languages. (Graham and Zook, 2013) A significant problem with this method lies in the univariate design of the visualization. The authors were not able to represent polylingualism with this visualization. Such a design is not possible without corroborating multiple language versions of pages with their geospatial locations, a

that would require cross-referencing thousands of webpages to find spatial and linguistic equivalents.

Another method employed in the collection of linguistic diversity is to collect linguistic instances from the linguistic landscape.

2.3 GIS and the Linguistic Landscape

A oft-cited definition for linguistic landscape comes from *Linguistic Landscape and Ethnolinguistic Vitality*, the seminal work by Landry and Bourhis, stating it “refers to the visibility and salience of languages on public and commercial signs in a given territory or region” (Bourhis, Landry, 1997:23). It is the last six words of this definition that suggest a spatial function to linguistic landscape, and that in turn, linguistic landscape may serve as a viable spatial indicator requiring the toolsets of GIS analysis. Further, the “informational function of [LL] ...serves to delineate the territorial limits of the language group it harbors relative to other linguistic communities inhabiting adjoining territories.” (Bourhis, Landry, 1997:25)

In applying geographic analysis to the linguistic landscape, there are natural concerns raised over the quantitative aspect of language unit—namely, what constitutes the borders of a linguistic instance, i.e. a token of language. According to Backhaus, “any piece of text within a spatially definable frame” constitutes a single instance, from “small handwritten stickers to huge commercial billboards.” (Gorter, 2006:3) However, in the case of Cenoz and Gorter, the definition of a token for “shops, banks and other businesses to take all texts together as a whole and thus each establishment and not each individual sign became the unit of analysis.” (Cenoz, Gorter, 2007:71) Though it seems prudent to corroborate a collection of LL tokens with the property or cadaster, if a purely linguistic approach does not intend to use such data, then it is superfluous and should not be collected in the data collection. Thus, Backhaus’s definition presents a more appropriate approach to collecting linguistic tokens for a study of multilingualism.

Language policy may greatly influence or even dictate language presence in a geographic area, from making recognition of a minority language as official policy to providing incentives for the inclusion of a dominant language. In the former case, the state of Ethiopia enacted constitutional measures that guaranteed “persons belonging to various ethnic and linguistic minorities shall not be denied the right to enjoy their own culture and to use their own language.” (Woldemariam, Lanza, 2014:82) Such dictates do much to fortify the linguistic landscape, especially considering the fact that other written forms of Ethiopian languages were forbidden up until 1976. Conversely, the state can do much to enforce the use of the dominant language. In the case of Bangkok, to encourage the use of the national language, “the government provides a tax incentive for including Thai on commercial signs.” (Huebner, 2008) Also noted in by Huebner is an official exceptions to these incentives, such is the case where official minority signs are used to promote an ethnic minority village as a tourist destination. Backhaus notes rather vaguely that in Tokyo, “official multilingual signs without Japanese are a rare sight.” (Backhaus, 2008:57) In the case of Israel, the positions of language policy can be seem to be conflicted. Many official signs are trilingual Arabic-English-Hebrew, while official policy.

“Russian, as well as other immigrant languages, are generally ignored even when a clear claim for cultural and linguistic recognition. As a rule, top-down LL items in Jewish localities tend to ignore immigrant languages and to make do with Hebrew and English. [LL] is not faithfully representative of the linguistic repertoire typical of Israel’s ethno-linguistic diversity, but rather of those linguistic resources that individuals and institutions make use of in the public sphere.” (Ben-Raphael et al, 2008)

In the case of Malaysia, language policy has guaranteed the rights to ethnic and linguistic minorities, in much the same way as Ethiopia. In fact, it is due to Malaysia’s educational system that so many ethnic minorities are largely polylingual.

The majority of the current generation of Malaysian Chinese youth go through Mandarin primary education and Malay-medium secondary schooling in the Malaysian public education system. Hence, an average Malaysian Chinese youth knows, at varying levels of proficiency, at least three languages namely, Mandarin, Malay and English. (Ting, 2013)

Conversely, some recent Malaysian language policy measures can be seen as directly affecting the linguistic landscape for the benefit of a single group.

The Malacca [State] government’s instruction that the Jawi script be used on signboards of premises and billboards should be followed by others so as to preserve the script which has been a part of the identity of the Malays. (Malay Mail, January 11, 2016)

When analyzing the linguistic landscape, one is presented with a set of linguistic configurations assembled from various combinations of languages scripts present in an area. In decomposing the dataset, these configurations offer key insights into the state of multilingualism in that area. Greater still may be the quantity of key scripts and co-occurrences of script in telling the state of ‘power and agency’ in an area. Huebner notes:

...variation may also be a reflection of the relative power and social status of various groups within a given community, and/or the nature of the activities these groups are engaged in. (Huebner, 2008)

In his analysis of Bangkok signs, Huebner decomposes signs by multi-linguistic configuration as well as by official or non-official signs, or ‘*in vitro*’ and ‘*in vivo*’, terms coined by Calvet (Calvet, 1990). While indicating these criteria may be key in understanding the role of government and power in the linguistic landscape, it also helps to understand the growth of status roles in language, something that is largely influenced economic forces. Indeed, Raphael developed, and later Cenoz and Gorter applied collection scheme with no fewer than sixteen criteria per single linguistic landscape token. (Gorter, 2008, from Cenoz, Gorter) However, a geographic analysis constrained to the spatial distribution does not require additional dimensions of the linguistic landscape to be collected; only quantity and geography are key. Hence, collection of multilingualism should reflect language ‘as it is experienced in the environment’ without other ‘qualitative distinction[s]’. (Backhaus, 2008)

M E T H O D O L O G Y

3.1 Overview

The objective of this thesis is to use SANET kernel density estimation to analyze the co-location patterns of multilingual instances of *lingua francas* in Malaysia's linguistic landscape over a selection of regions and scales using stacked line graphs to visualize weighted local aggregations and custom linguistic diversity indicators in analogous study regions and scales. The use of kernel density estimation can be understood as the natural first response tool to the following questions elicited of the linguistic landscape.

What is the linguistic landscape?

Landry and Bourhis claimed the linguistic landscape “refers to the visibility and salience of languages on public and commercial signs in a given territory or region” (Bourhis, Landry, 1997:23). It is the last six words of this definition that suggest a spatial function to linguistic landscape, and that in turn, linguistic landscape may serve as a viable spatial indicator requiring the toolsets of GIS analysis. Further, the “informational function of [LL] ...serves to delineate the territorial limits of the language group it harbors relative to other linguistic communities inhabiting adjoining territories.” (Bourhis, Landry, 1997:25)

Throughout peninsular Malaysia, the linguistic landscape offers vast tapestries of street-level photographic coverage of Malay, English, Tamil, Chinese scripts, and Arab-Jawi scripts, clearly present in mono-, bi-, and polylingual configurations, now freely available for high-volume data collection through supervised geo-tagging. Malaysia presents an intriguing combination of ‘inclusive’ and ‘exclusive’ scripts at a high national linguistic diversity measure of .758 (UNESCO, 2009). A quantitative analysis of the linguistic landscape requires a GIS system to perform such an analysis.

What is the indicator design?

A few notable linguistic studies employing GIS to analyze multilingualism have used census data in their indicator design [(Luebbering, et al (2013), Veselinov, (2009), Van der Merwe (1993)] These measures have their roots in Greenberg's linguistic diversity measure. (Greenberg, 1956) Other studies have included aggregate internet search results in their indicator design for multilingualism. (Liao and Petzold (2010), Graham and Zook (2013) Pimenta, Prado, Blanco, 2009) Using the definition of a linguistic token defined by Cenoz and Gorter, (2007), the linguistic landscape indicator first decomposes aggregate data by language (script) and multilingualistic conglomeration. This design further suggests a number of non-spatial metrics. These non-spatial metrics may suggest insights regarding the ‘power and status’ relationships in the linguistic landscape. (Calvet, 1990, Woldemariam, Lanza (2014), Cenoz, Gorter (2007), Bourhis, Landry (1997) These customized metrics include (η) *Monolingual Composition*, (C) *Bilingual Co-occurrence Unions* (e.g. #BM \cap #BT), (A) *Affiliation [C/(T- η)]*, *Lingua Franca Preference*, and Greenberg's Linguistic Diversity Index. (See 3.2 Multilingualism Indicator Design).

To whom is a linguistic landscape indicator of interest?

Though not a complete list, the linguistic landscape indicator is expected to be of great interest to the relationship for those in education and educational policy (Cartwright,

D.(2006), Cenoz, Jasone; Gorter, Durk (2008), Gorter, D., & Cenoz, J. (2007), Torkington, K. 2009, Cenoz and Gorter (2008), and Gorter, D., Marten, H. F., & Van Mensel, L. (Eds.) (2012); emergency management; demographers; the intelligence community; election data and campaign administrators such as undi.info; linguistics projects such as the LL-MAP project. It is expected this spatial examination of the linguistic landscape will be of particular interest and a contribution to the discussion of current mapping trends in linguistic landscape mapping discussed by Barni and Bagna (2009).

How is the data gathered?

Though the use of *Street View* has been suggested by Barni and Bagna (2009) and later by Inoue (2012), there is an apparent disparity in the literature in which *Street View* is implemented as a tool for quantitative linguistic landscape data collection practices intended for spatial analysis by GIS. Google *Street View* is the exclusive data collection method employed in this analysis and intended to be of great value in further linguistic landscape analysis. The linguistic landscape tokens are geo-tagged in Google Earth and the KML points are map matched and analyzed in SANET network toolkit with kernel density estimation.

What kind of sampling method should be used?

The sampling method employed for data collection is the total population sampling method. Total population method is a complete sampling of the population which, applied to data collection of the LL, samples an entire population sharing the same geographic location. This purposive sampling technique will make “it possible to get deep insights into the phenomenon you are interested in... make it possible to make analytical generalizations about the population being studied.” (Laerd, 2017)

A 110-km long corridor running from the heart of the tourist area in urban Melaka to the FELDA palm oil communities in the middle of the peninsula. While this method of using a single network path greatly reduces local precision and may affect the functionality of the SANET tools, it is assumed that the corridor used in this study represents only a single ‘edge’ in what could be part of a great reticulated network of linguistic landscape corridors. This sacrifice is made in order to visualize global trends.

What are ‘along-network’ points?

Points in a network are considered to be events that occur on a road surface or directly above it, according to the SANET manual (SANET: Spatial Analysis Along Networks, 2015). The linguistic landscape consists mostly of along-network points. These points may sometimes be visible and legible many meters from the network edge to which they are adhered or *matched*. Linguistic landscape occurs along transportation networks and, because of this, is not suitable for Euclidean (planar) analysis methods but instead should be analyzed by linear network methods.

How are these points matched?

After the linguistic landscape tokens are collected in the business nuclei of several small towns and along the single main corridor connecting these small town and one large

urban area, the points along this 110-km corridor are matched onto a single network path using a technique called 'Mainstreeting'. This includes matching tokens taken from parallel and intersecting streets. 'Mainstreeting' is a kind of local spatial aggregation that allows for the visualization of co-location patterns. 'Mainstreeting' is a method of collecting tokens along the side streets in the small towns in order to have a more complete and total picture of the linguistic make-up of the near-along points by concentrating total local values upon a single corridor. These points are believed to not require weighting as long as they are not matched in excess of the bandwidth measure.

Why are kernel density estimation measures used?

This first moment analytical tool is best suited to find mean-centers and direction of orientation in the linguistic landscape trends. KDE is more suitable as the initial tool to study the regional non-urban trends in the linguistic landscape than Hot Spot Analysis. Hot Spot Analysis may be suitable for a mean-adjusted sample or for a single urban area. However, there is currently no Getis-Ord G_i^* or Hot Spot Analysis tool in the SANET toolkit. The advantage in using kernel density estimation lies in being able to adjust the bandwidth in order to use the most suitable radius for the trend. KDE is also the only suitable to visualize co-location patterns as a 3-D projection.

What bandwidths are used?

The kernel density estimations bandwidths are performed at 10-m, 100-m, and 500-m. The intention is to reveal trends at the municipal (10 m), voting and school district, and state level (500-m). Using these three bandwidths should effectively reveal local and global geographic trends at these scales. Cell size is kept at 1000 m.

Why are K-functions suggested?

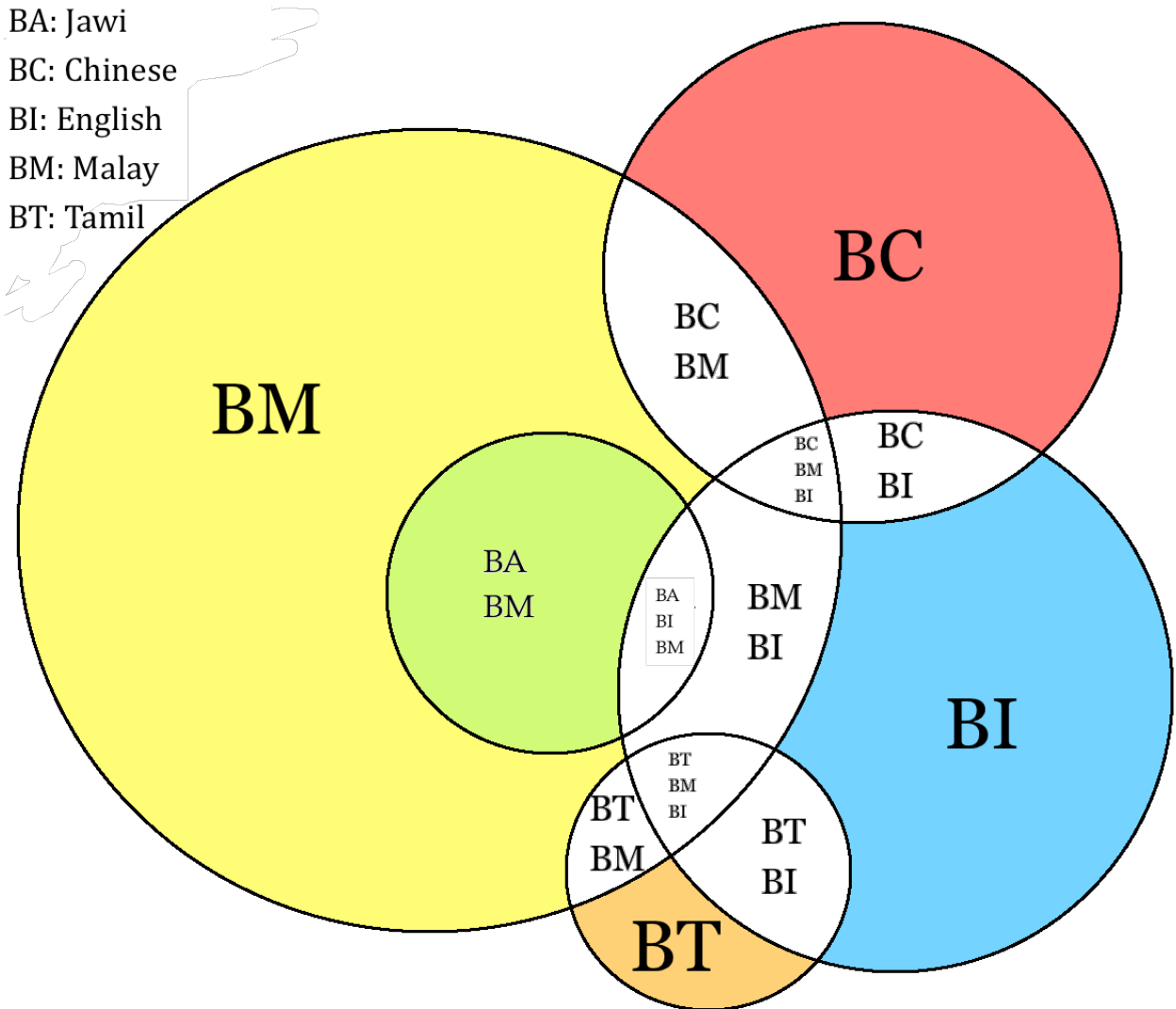
A K -function will give statistical significance to the distribution of linguistic sets, allowing for probabilistic interpolations of clustering and dispersion at a range of scales. K -functions offer a insights into the second moment trends that occur in the variance. While first moment trends can allow us to see mean centers and direction of orientation, second moment K -functions offer probabilistic expectations of an event occurring at an r distance from an event. (Smith, 2011) However, implementing a use of K -functions on the linguistic landscape is beyond the scope of this study.

How are the results visualized?

Co-location patterns are able to be visualized along the 110-km corridor by double or triple projecting the density measures along the same corridor. These results are then extruded and the transparency is adjusted so that the result is a stacked line graph. Though by no means a perfect method, visualizing the results using the extrusion and transparency methods intuitively allow for a good understanding of both the trend and the relative difference in linguistic set presence. Absolute values may be indicated with line measures. The map type produced by this method has been designated a 'trend map'.

3.2 Multilingualism Indicator Design

In the Malaysian vernacular educational system, “an average Malaysian Chinese youth knows, at varying levels of proficiency, at least three languages, namely: Mandarin, Malay and English.” (Hung:2013:83) This relationship can be seen as an L1 Chinese speaker’s relationship with an L2 and L3. The figure titled *AGGLOMERATION OF MULTILINGUAL RELATIONSHIPS* illustrates the multilingual relationships between ethnic L1 groups and of the current educational system in Malaysia:



AGGLOMERATION OF MULTILINGUAL RELATIONSHIPS

The colored regions indicate an L1 group; the unions represent L2 and L3 multilingual interactions of L1 with an L2 and L3. This linguistic agglomeration consists of four languages and one script. Two of these languages are inclusive languages: Malay, abbreviated as BM (Bahasa Melayu), and English, abbreviated as BI (Bahasa Ingerris). The remaining three regions are an L1 spoken or read almost exclusively by an ethnic group: Chinese dialects [BC(Bahasa Cina)], Tamil [BT(Bahasa Tamil)], and Jawi, which is the Malay

language written in the Arabic script [BA(Bahasa Arab)]. Of the two *lingua francas*, Malay is spoken by an ethnic group of whom some use an exclusive script, Jawi; English has no ethnic group affiliation and thus a member of any Malaysian ethnic group can claim English as their L1.

The spatial distributions of *lingua franca* are of particular interest to this research and the main aspect of linguistic diversity in this research. A quote by Greenberg best describes this motivation:

Our general expectation, subject to significant qualifications, is that areas of high linguistic diversity will be those in which communication is poor, and that the increase of communication that goes with greater economic productivity and more extensive political organization will lead typically to the spread of a lingua franca. (Greenberg, 1956:110)

Multilingual instances may include the presence of one lingua franca (Malay or English) more than another, or both; and that preference occur with a with observable spatial distributions. The significance of *lingua franca* and *Lingua Franca Preference* is presumed to be best tested by analyzing the co-occurrence of BC::BI unions against the BC::BM unions, exclusive of and inclusive of monolingual tokens in each respective territory. It is the intention of this thesis to demonstrate that the local and global distributions of Malaysian *lingua franca* preference can be measured using first moment and second moment analytical methods, namely kernel density estimation and *K*-function. With this in mind, the indicator design is the next step.

Conceptual Design

As a linguistic indicator, numerous policy-making decisions can be made, from influencing language policy to educational reform; from to emergency response management. Such an indicator could be critical in the management of a PSA campaigns in high risk vectors for Dengue. In the figure below, two public service announcements warn of the risk of Dengue fever. Note two PSA banners for the risk below, one in Malay and another in Chinese:



© COURTESY OF GOOGLE STREET VIEW: TWO VERSIONS OF A DENGUE FEVER PSA

Linguistic Landscape Indicator Design

Resources	Metrics	Products	Beneficiaries
<p>University of Malaya</p> <p>Linguistics Department researchers may be able to provide in depth knowledge of not only historical language shifts but also geographical changes. Department may also be invaluable in interpreting complicated loanword issues for tagging convention.</p>	<p>Linguistic Landscape Aggregate Diversity</p> <p>All tokens on a corridor decomposed into sets of instances. Aggregate statistics will indicate:</p> <ol style="list-style-type: none"> 1. Total Number of Instances 2. Percent of Aggregate 	<p>LL-Malaysia Portal</p> <p>An Java-based application developed for GIS visualisations of Malaysia's linguistic landscape at global and local resolutions using historical Google StreetView coverage.</p>	<p>Ministry of Education</p> <p>From local <i>guru besar</i> to Ministry of Education, administrators and policy makers would benefit greatly from informed the decision-making capacity made possible by knowing the linguistic composition of a desired region.</p>
<p>Heuriscapes, LLC</p> <p>A foreign GIS project management firm specialising in GIS for education and designing analytics for education. This firm may most successfully navigate both the linguistics, educational, and technical requirements needed for such a project.</p>	<p>(η) Monolingual Composition</p> <p>The monolingual composition measure is:</p> <ol style="list-style-type: none"> 1. A total number of monolingual instances of out of all single language content; 2. A total number of monolingual instances of out of all single language content out of all content. 	<p>Map Products</p> <p>Kernel Density Estimation using both planar and non-planar calculations. <i>K</i>-function calculations using both planar and non-planar methods. Local and global calculations.</p>	<p>Emergency Management</p> <p>With the spread of vector-borne diseases, especially Dengue fever, it is necessary to execute effective PSA campaigns in order to effectively communicate. This also includes other disaster preparedness measures such as flood escape route planning.</p>
<p>Putra Malacca Development, Sdn. Bhd.</p> <p>An application developer chosen to write the code and develop and host the Java-based application.</p>	<p>(<i>C</i>) Bilingual Co-occurrence</p> <p>Co-occurrence Unions, such as #BM∩#BT, measures bi- and multilingual co-occurrences of linguistic unions.</p>		<p>Demographers</p> <p>Though SANET forbids the use of its GIS application for commercial purposes, there are non-commercial applications of interest to demographic studies, especially those of religious affiliation. Local and regional offices of commerce may benefit from profiling their geography.</p>
<p>VGI</p> <p>Using the efforts of independent contractors and amateur linguists to maximise the extent and minimise the costs of a linguistic landscape dataset. May includes interior views of buildings, properties, and other private areas for the soul purpose of anonymous linguistic metadata.</p>	<p>(<i>A</i>) Affiliation [<i>C</i>/<i>T</i>-η]</p> <p>Affiliation is a metric designed to measure the influence of one language upon another. Results offered in a matrix with values ranging from 0 to 1 from least to greatest.</p>		<p>Intelligence Community</p> <p>[REDACTED]</p>
	<p>(<i>P</i>η) Preference</p> <p>The Preference measure indicates the influence exerted by a <i>lingua franca</i> by an aggregate monolingual quotient:</p> <p>$[\eta\#BI/\eta\#BM]$ by a bilingual quotient: $[[\{BA\cap BI\} + \{BC\cap BI\} + \{BI\cap BI\}]] /$ $[[\{BA\cap BM\} + \{BC\cap BM\} + \{BI\cap BT\}]]$ by a trilingual quotient: $[[\{BA\cap BC\cap BI\} + \{BA\cap BI\cap BT\} + \{BI\cap BC\cap BT\}]] /$ $[[\{BA\cap BC\cap BM\} + \{BA\cap BI\cap BT\} + \{BM\cap BC\cap BT\}]]$ and even quadrilingual quotient: $\{BA\cap BC\cap BI\cap BT\} / \{BA\cap BC\cap BM\cap BT\}$</p>		<p>UNDI</p> <p>Election data and campaign administrators may benefit greatly from understanding the linguistic terrain. Organisers may find it an invaluable tool in distributing campaign efforts and plan media events.</p>
			<p>LL-MAP Project</p> <p>"An [academic] collaboration between linguists, historians, archaeologists, ethnographers, and geneticists, ...[exploring] the relationship between language and cultural adaptation and change." (LL-MAP.org, 2017)</p>

3.3 Data Collection Method

In applying geographic analysis to the linguistic landscape, there are natural concerns raised over the quantitative aspect language units—namely, ‘What constitutes the borders of a linguistic instance, i.e. a token of language?’ According to Backhaus, “any piece of text within a spatially definable frame” constitutes a single instance, from “small handwritten stickers to huge commercial billboards.” (Gorter, 2006:3) However, in the case of Cenoz and Gorter, the definition of a token defines the property on which it is resident, including “shops, banks and other businesses to take all texts together as a whole and thus each establishment and not each individual sign became the unit of analysis.” (Cenoz, Gorter, 2007:71) Backhaus’s definition presents a more appropriate approach to collecting linguistic tokens and will be used for this study of multilingualism.

The sampling method employed for data collection is total population sampling method. Total population method is a complete sampling of the population which, applied to data collection of the LL, samples an entire population sharing the same geographic location. This purposive sampling technique will make “it possible to get deep insights into the phenomenon you are interested in... make it possible to make analytical generalizations about the population being studied.” (Laerd, 2017)

Collecting cadastral data may be quite valuable if and only if linguistic data is not embedded in a single property agglomeration, and though corroborating a collection of tokens with a property or cadaster can easily be done, this property assignment will not be done in this study of multilingualism. All tokens will be treated as distinct instances of written language. Caution should be used with such a practice to prevent distorted concatenations from occurring. At least one highway used in the sample in this study that is itself an administrative division. This would require splitting linguistically distinct edges, dividing linguistically similar street from their opposite side and aggregating them with dissimilar ones. Conversely, one may keep the similar edges intact and visualize the street segment as a *dual graph*, though it would lose the native shape of administrative polygons, and thus some degree of GIS computational functionality.(Boeing, 2016)

The following images depict some typical imagery seen in the Malaysian linguistic landscape:

A. BOOK STORE, KUALA PILAH (#BC, #BI)



B. GOLD SHOP, KUALA PILAH (#BA, #BM)



C. CLINIC, KUALA PILAH (#BC, #BM, #BT)



D. PRODUCE STAND, KUALA PILAH (#BC, #BM)



E. JEWELRY SHOP, KUALA PILAH (#BA, #BC, #BI, #BM, #BT)



F. RESTAURANT, KUALA PILAH (#BM, #BT)



G. MINI MART, TAMPIN (#BA, #BC, #BM, #BT)



H. HINDU TEMPLE, TAMPIN



I. EYEGLASSES AND CONTACTS SHOP, MASJID TANAH



J. HOLIDAY GREETINGS FROM STATE GOVERNMENT, TAMPIN



K. CLINIC, KUALA PILAH (#BA, #BC, #BM, #BT)



L. BARBER SHOP, TAMPIN (#BA, #BC, #BM, #BT)



M. TAILORING SHOP, KUALA PILAH (#BC, #BI)



One issue that presents itself in the process of data collection is the preponderance of loanwords that exist in the Malay language. . In L., the term ‘Boss’ is used with an identical spelling in both Malay and English. Though the spelling of these loanwords often differs from the English or other foreign orthography, as is the case of token F. and J., note the spelling of ‘restoran’ and ‘Krismas’ respectively, in some cases the sign as a whole must. For example, in A. and M., the use of Chinese names in pinyin have been combined with English, as seen in A. ‘Boon Hwa Book Store’

Total population sampling is the sampling method used in this study due to the token size per road segment, the significance of measuring near-total linguistic variation per road segment, and employing Google *Street View* to its full capacity. Namely, sampling to the extent to which the photographic resolution allows possible.

Total population sampling is a type of purposive sampling technique that involves examining the entire population (i.e., the total population) that have a particular set of characteristics (e.g., specific attributes/traits, experience, knowledge, skills, exposure to an event, etc.). (Research Methodology, 2017)

Total population sampling, as implemented by this definition, is the practice of sampling all linguistic tokens visible and legible within the Google *Street View* coverage. It should be noted that some road segments of Google *Street View* coverage differ greatly from others In their resolution due to any number of factors: weather conditions, time of day, dirt occlusion, angle of picture, or even camera model. Future LL data collection efforts may find it useful to include the photographic resolution of a segment as a weighting element.

Due to the basic design of the dataset, the only the criteria collected from the signs are the language metadata. The naming convention tagged all tokens with a ‘#’ symbol

followed by the Malaysian language abbreviation for the language. The simplest means of extending data collection to include further criteria is to input data separated by commas in order to be read as a CSV. In the case that additional linguistic landscape criteria are to be collected, a good CSV compatible convention be to use a concatenated identifier. Instead of using commas to indicate a #BC, #BI instance, a concatenated identifier could be written as following: (BA0BC1BI1BM1BT0). Such a practice of using comma-separated values keeps the KML code shorter and allows it to be more easily extracted from the KML as all fields are resident in a single line.

The sampling area is taken from four states in peninsular Malaysia—Melaka, Pahang, Kelantan, and Negeri Sembilan. The bulk of the sample is a continuous route that runs from the urban area of Melaka, runs through Sungai Udang, Masjid Tanah, Tampin, Kuala Pilah, terminating in Bahau, Jempol, Negeri Sembilan. Several dis-continuous areas have been sampled: Seremban Town, Palong Felde 7/8, Jempol, and Tanah Merah, Kelantan.

3.4 Kernel density estimation

In order to detect trends in the linguistic landscape we must select the most suitable analytical tool or tools. Both first moment density measures and second moment variance functions can offer key insights into the patterns of linguistic landscape data.

Kernel density estimation “is one of the most popular methods for analyzing the first order properties of a point event distribution.” (Bailey & Gatrell, 1995; Silverman, 1986 from Xie, Yan, 2008) As a first moment mean-smoothing analytical tool, it “focuses on the underlying properties of point events and measures the variation in the mean value of the process”. (Xie, Yan, 2008:396) The great advantage of kernel density estimation is by aggregating mean values over a distance by ‘mean smoothing’ with the user’s ability to adjust the bandwidth. This adjustment achieves a desired ‘local intensity’ by mean smoothing over a one-dimensional space(Diggle, 1985). The search bandwidth r determines the amount of kernel smoothing according to the search radius (r). The greater the bandwidth, the greater the amount of mean-adjusted smoothing, while a smaller search radius r results in local clustering of hot spots.

A formula for planar KDE from Xie and Yan (2008) can be seen here:

$$\lambda(s) = \sum_{i=1}^n \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right)$$

“where $k(s)$ is the density at location s , r is the search radius (bandwidth) of the KDE (only points within r are used to estimate $k(s)$), k is the weight of a point i at distance d_{is} to location s . k is usually modeled as a function (called kernel function) of the ratio between d_{is} and r .” (Xie, Yan, 2008: 397)

Having the ability to adjust r is critical in making scalar observations of linguistic landscape data. The clustering of linguistic landscape configurations—often the multi-linguistic configuration that include the language or script of at least one ethnic minority—usually occurs at a number of scales of interest: at several blocks of a main street, at the

neighborhood level in urban areas, at the district and state level. Using a small bandwidth radius will elicit ‘spikes’ when detecting a high clustering density. Additionally, it is also necessary to measure more global trends for anisotropy in linguistic data. When these trends occur, the tools should not detect clustering in the linguistic landscape with such sensitivity, but utilize a measure that accounts for the scale at which a trend occurs. This entails the use of larger bandwidth. Kernel smoothing would be appropriate for detecting regional trends in *lingua franca* preference; district and state level measures may require a greater degree of smoothing to come closer to detecting trends of anisotropy. Indeed, it would seem the tool produces viable insights into trends in linguistic landscape at multiple scales. However, there is a problem with planar kernel density estimation producing biased results when used on events that occur in a network. These can include traffic accidents or along-network events, such as linguistic landscape data. The problem is that planar KDE bias has been known to over-detect clustering near nodes in a network when a non-Euclidean measure is more appropriate.

“estimation produces a bias, because when points are distributed according to a uniform distribution on a network, the density estimated by the above method does not produce a uniform distribution.” (Okabe, Satoh, Sugihara, 2009: 8)

Such a problem becomes more significant in urban areas, where there are a higher number of nodes intersect and cause a greater number of bias. Such a problem is avoided when are calculated in a constrained network space rather than in Euclidean space.

Thus, the first moment method of kernel density estimation is suitable for this analysis granted it can be conducted on LL data in a one-dimensional network space. Additionally, a second moment method is chosen for its theoretical value but is not implemented in the application.

Nearest Neighbor Analysis

In order to detect trends in the linguistic landscape, one may use clustering and dispersion methods of second moment events. One such method to detect clustering and dispersion is to use computed *Z scores* and *R scores* for NNA (Nearest Neighbor Analysis). NNA first computes the mean *R*, which then calculates a spatial mean center, followed by the second moment, the calculation of variance. By definition, “A numeric description of how values in a distribution vary or deviate from the mean.” (ESRI 2016)

NNA is sensitive to both scale and extent of the study area, as “it is important to note that we need to be careful in selecting an appropriate geographic scale to properly display objects and delineate the study area.” (Myint, 2008:176) However, there is no way to analyze multiple scales using NNA without some form of re-aggregation. (Ripley, 1996:§6.2) Since an LL dataset necessarily requires analysis at a range of scales—local, state, and regional— LL analysis should favor a method capable giving scalar results.

Ripley’s K-function

The *K* -function is probably one of the most widely used methods in many spatial analysis approaches dealing with point distribution on a plane (Myint, 2008:176). The simplest use of Ripley’s *K* function is to test Complete Spatial Randomness, i.e. to test

whether the *observed* events, the LL points, are consistent with a homogeneous Poisson process, an *expected* event field exhibiting CSR and assumed isotropic, with K being a quotient of the two across a range spectrum. This spectrum indicates clustering and dispersion across a spectrum of increasing r .

For a set of points, Ripley's K can indicate clustering or dispersion of a point set against a theoretically random event. The first moment property of a spatial point pattern is the number of points per area and the second moment property is the expected number of points N within a distance r of another point. Ripley's K -function is the second moment property normalized by the density (or intensity), the number of points per area λ . (Kiskowski, Hancock and Kenworthy, 2009:1095)

The advantage of Ripley's K lies in its use across a range, and it is necessary to obtain results at a range of scales for the LL dataset.

The advantage of K -function analysis is that it uses all point-point distances, not just the nearest neighbor distances, to show spatial clustering at various scales of pattern, and the distance where clustering or over-dispersal becomes significant. (Bailey and Gatrell 1995)

Ripley's K is useful at finding trends in the linguistic landscape, but it is necessary to determine the significant scale ranges. Theoretically, dispersion should be significant above the five meter distance r —the width of a shop front. The spacing intervals of property and cadastral units are clearly a factor that would point to false dispersion at this range. Clustering may be detected at the 20-m meter range, roughly the width of a small town main street; however, the detection of 'Main Street' clustering depends on using points prior to map matching of points. Urban areas, with normalized grids of streets, might see dispersion at the block r distance of 45-m and 80-m scale range due to the incremental spacing.

Another significant scale is the district level scale: one that is large enough to bisect a small city with a spatial extent of several kilometers. A 1000-m to 5000-m r may be large enough to reveal multilingualism clustering in neighborhoods, commercial areas, and areas of specialized industry, i.e. automotive. Further still is the degree to which *global* K -functions reveal multilingual correlations across regions at the 10-km to 50-km r . The clustering of a script seen in one region may show high dispersion in other parts of the country. This may contrast greatly with the homogenous areas of dominant script of another region.

In semi-rural areas between small towns connected by a single highway, a network with a high average street (edge) length and few nodes can approach a one-dimensional graph. Because the linguistic landscape datasets occur along a network, there may be vast reticulated areas of empty space occurring in the study area. This non-Euclidean spatial constraint will adversely affect the measure, possibly indicating false clustering.

When used to analyze spatial point patterns constrained by road networks, the K -function can result in over-detection of clustering patterns, leading to possible Type 1 errors. (Yamada and Thill, 2003:149).

3.5 SANET

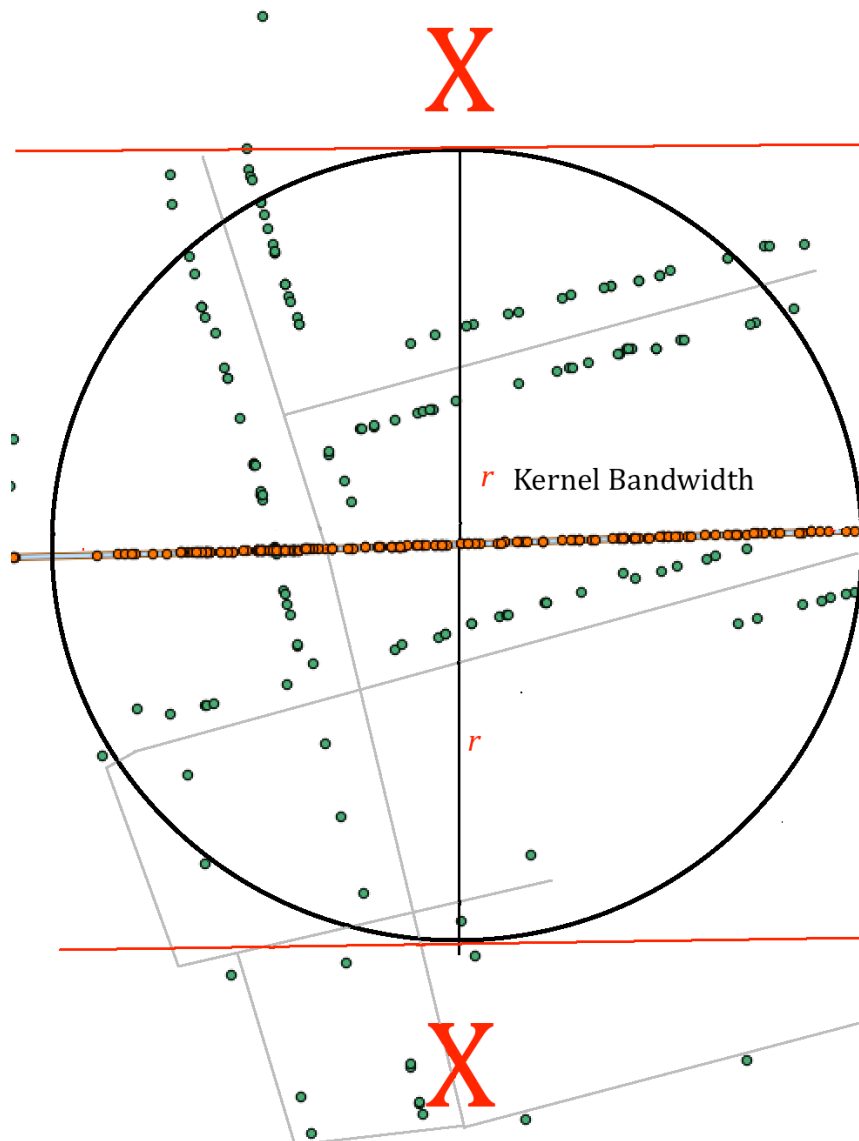
SANET kernel density estimation satisfies the requirements needed for linguistic landscape analysis in order “to analyze spatial phenomena that occur on networks...both [on-network] and [along-network.]” (Okabe et al, 2006:57) The computation of network kernel density analysis differs from the planar computation in one significant way: Instead of computing over an areal unit, it computes over a linear one, or combination of linear paths (Xie, Yan, 2008). The formula for linear computation of kernel density estimation can be seen in the following:

$$\lambda(s) = \sum_{i=1}^n \frac{1}{r} k\left(\frac{d_{is}}{r}\right)$$

“There exists a consensus that the choice of the kernel function k is less important than the choice of search bandwidth r .” (Bailey & Gatrell, 1995; O’Sullivan and Unwin, 2002; Schabenberger & Gotway, 2005; Silverman, 1986; O’ Sullivan and Wong, 2007). Kernel density estimation methods that use a univariate kernel method have been shown to lead to edge problems. Solutions to these edge problems have been proposed in the literature [(Tapia and Thompson 1978, Devroye and Gyorfı 1985, Silverman 1986, Scott 1992, Wand and Jones 1995, Devroye and Lugosi 2000, Eggermont and LaRiccia 2001) from Okabe, Satoh, and Sugihara, 2011:11]; however, SANET avoids this edge bias through equal-split kernel functions in which kernel density estimation methods can calculate bandwidths as both continuous and discontinuous paths at the node(Okabe, Satoh, Sugihara, 2009).

The kernel density estimation implemented in this study employs a simplified network design in which there are no intersections. The network is constructed without intersections for a number of reasons. A single linear path that matches along-network points to a single non-intersecting network line makes it exceedingly easy to visualize multiple linguistic query expressions as one continuous stacked line graph. Such a technique has a tradeoff: it sacrifices precision at finer scales in order to retain cartographic coherency. This technique is not intended for use on the linguistic landscape of a single urban area at fine scales. When visualizing regional density measures, stacking graphs allow for quantitative co-location as well as visual comparisons of linguistic populations. This technique is suitable for use only when the study area is large. Indeed, the largest sample corridor used in this study is a 110-km multi-state sample, and not a single urban area.

The following figures illustrate this technique:



'MAINSTREETING' WEIGHTING THRESHOLD

Aside from coherency, the desired effect of 'Mainstreeting' is to utilize a single network corridor to achieve the effect of 'local aggregation' in small towns, exurban, and suburban areas where the bulk of multilingual tokens are present. In effect, 'Mainstreeting' incorporates a larger sample area from secondary streets in an area and matches them to a single corridor in order to prevent occlusion during visualization. The trends in the linguistic landscape are examined on a 110 km corridor using bandwidths of 10 m, 100 m, and 1000 m. It is assumed that points may be aggregated without distortion along a single network line as long as the distance of the points are not matched in excess of the length of bandwidth r . It is assumed any point beyond the r threshold of the bandwidth must be distance weighted.

The SANET kernel density estimation tool output is two 3-D shapefiles: input points

and kernel density results as z-dimensional lines. Kernel density estimation parameters, namely bandwidth (r), can be adjusted in order to achieve a desired 'local intensity'. This mean smoothing process is performed on a 1-D point process. (Diggle, 1985:138)

SANET uses two functions, each differing in how they handle points at nodes. As Okabe (2009), 'equal-split continuous' and 'equal-split non-continuous'.

“differ in the assignment of values to l_2, l_3 and l_4 : the latter adjusts the values to make the function continuous in the 'local area' around the vertex v ; i.e., the area in which the distance from the vertex v is within $h-d$.” (Okabe, Satoh & Sugihara, 2009:19)

Because the network used for the 'Mainstreeting' method of concentrating local aggregation, and that this method is employed on a non-intersecting continuous network, there are no secondary nodes in which to equally split; no probabilistic paths l_2, l_3 or l_4 . So, in this case, 'Equal-split continuous' will be used for calculating the kernel density estimation measures along a single 1-D network path.

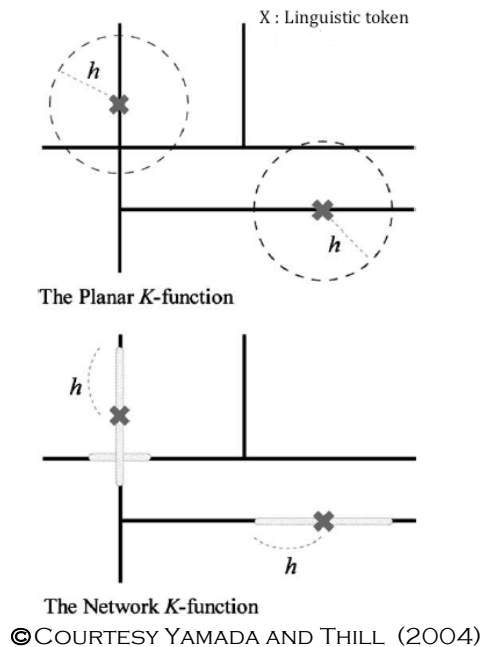
Network K-function

Though K -function analysis are not performed in this study, the theoretical groundwork is discussed in this methodology as the analysis is assumed to be of value in support of first moment mean density results; K -function and other variance analyses are the next step after GIS linguistic landscape mean density analysis.

It is a major assumption of this thesis that non-Euclidean measures of space should be employed on linguistic landscape data. At times referred to as Manhattan distance, network distance is measured on “a finite connected planar network consisting of a set of links and a set of nodes.” (Okabe and Yamada, 2001:272) As Ripley's K functions assume isotropy in all directions on a homogenous infinite plane with Euclidean distance, planar methods are inappropriate for use along-network datasets. Okabe and Yamada's (2001) model is represented as the following:

$$K(t) = \frac{1}{\rho} \mathbb{E} \left(\begin{array}{l} \text{the number of points } P \text{ within} \\ \text{network distance } t \text{ to a point } p_i \text{ of } P \end{array} \right),$$

Although “a method for computing a K -function on a network is much more complicated than the ordinary K -function method because a network is usually irregular”, a network K -function method has been developed by Okabe extended the K -function method to a finite network. (Okabe and Yamada, 2001:272)



The *K*-function is extended into network space by calculating the point density over shortest path distances using Dijkstra's (1959) shortest path algorithm. These shortest distance paths consider all possible routes from the observed point by constructing shortest path trees, and paths may change over the course of the calculation of t , the circle radius. (Okada and Sugihara, 2001:279-82) Because of this constrained network space, the network technique does not employ the same randomization techniques as *planar K*-function, namely, the use of Monte Carlo simulation.

In practice, Monte Carlo simulation is often carried out to produce pseudo-significance levels by repeated randomization. This technique determines the expected values of $K^{pl}(h)$ and the upper and lower significance envelopes under the null hypothesis of CSR. (Yamada and Thill, 2004:151)

The network *K*-function does not employ the Poisson process to generate a complete randomized space, but instead uses the binomial point process.

Since this is a binomial distribution, the stochastic point process of P where points of P are distributed according to the uniform distribution is called the binomial point process. (Ripley, 1981:255-66)

The binomial point process tests uniformity of the distribution of the observed points against their expected distribution over the finite linear network.

The assumption of the binomial point process is based on the hypothesis that points P are uniformly and independently distributed over a finite road network. Thus if this hypothesis is rejected, points P are spatially interacting and may form non-uniform patterns. (Spooner, Lunt, Okabe, and Shiode, 2003:493)

The randomization method is related to the edge effects. In the binomial point process, there are no edge effects because the network K -function assumes a finite space.

Edge effects can be caused either (i) by disregarding data outside the study region or (ii) by applying a statistic designed for an infinite space to a finite space. (Yamada and Thill, 2004:152)

Network cross K -function is the bivariate version of network Ripley's K . It employs the use of two kinds of points. There is no test for a CSR; the clustering and dispersion are calculated as interactions between the two point sets.

An assumption of linguistic landscape is that because samples are limited to the coverage of *Google Street View*, (or from roadside observations), which is almost exclusively composed of street-level photographic coverage of the LL, we cannot presume that the LL is a function of Euclidean space, but is instead a function of the transportation networks on which they are observed. Hence, as a function of transportation networks, non-Euclidean methods of analysis are used and SANET network analysis tools are employed.

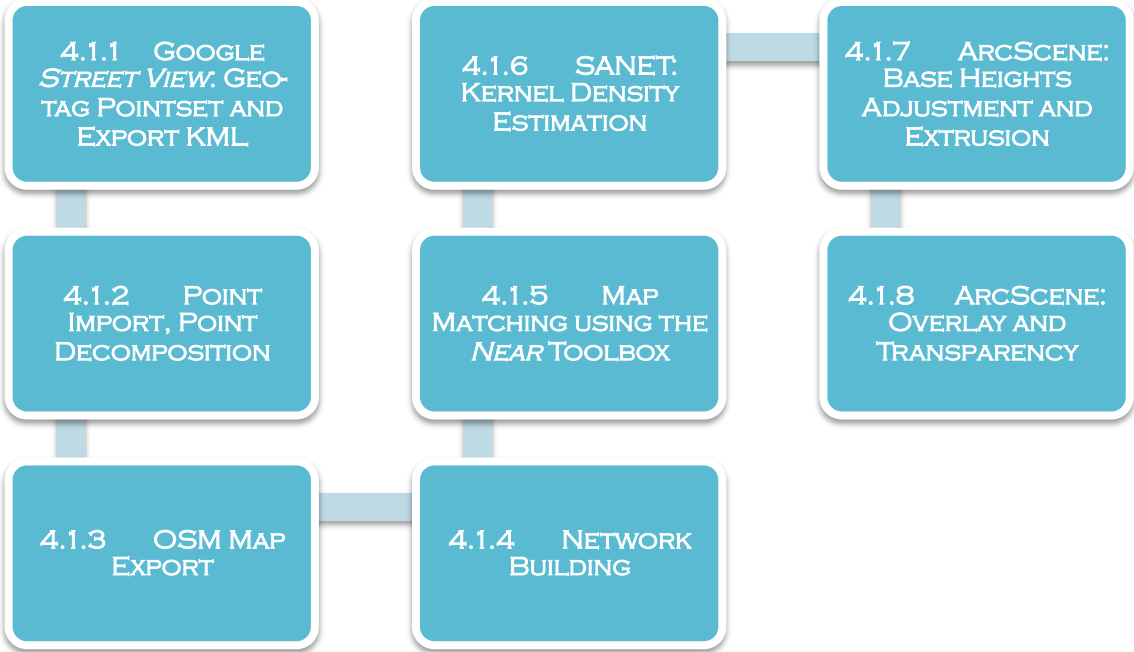
3.6 GIS Software and Resources

This study employs the use of QGIS interchangeably with ArcGIS for many functions. It also employs the use of network spatial analysis toolbox SANET 4.1 Standalone for all the network spatial analysis functions including kernel density estimation. *Google Street View* is used for data collection purposes. Additionally, *OpenStreetMap* is used for supporting the construction of polyline networks required by SANET analysis processes. Visualization is done exclusively in ArcScene for 3-D visualization of kernel density estimation shapefiles.

A P P L I C A T I O N

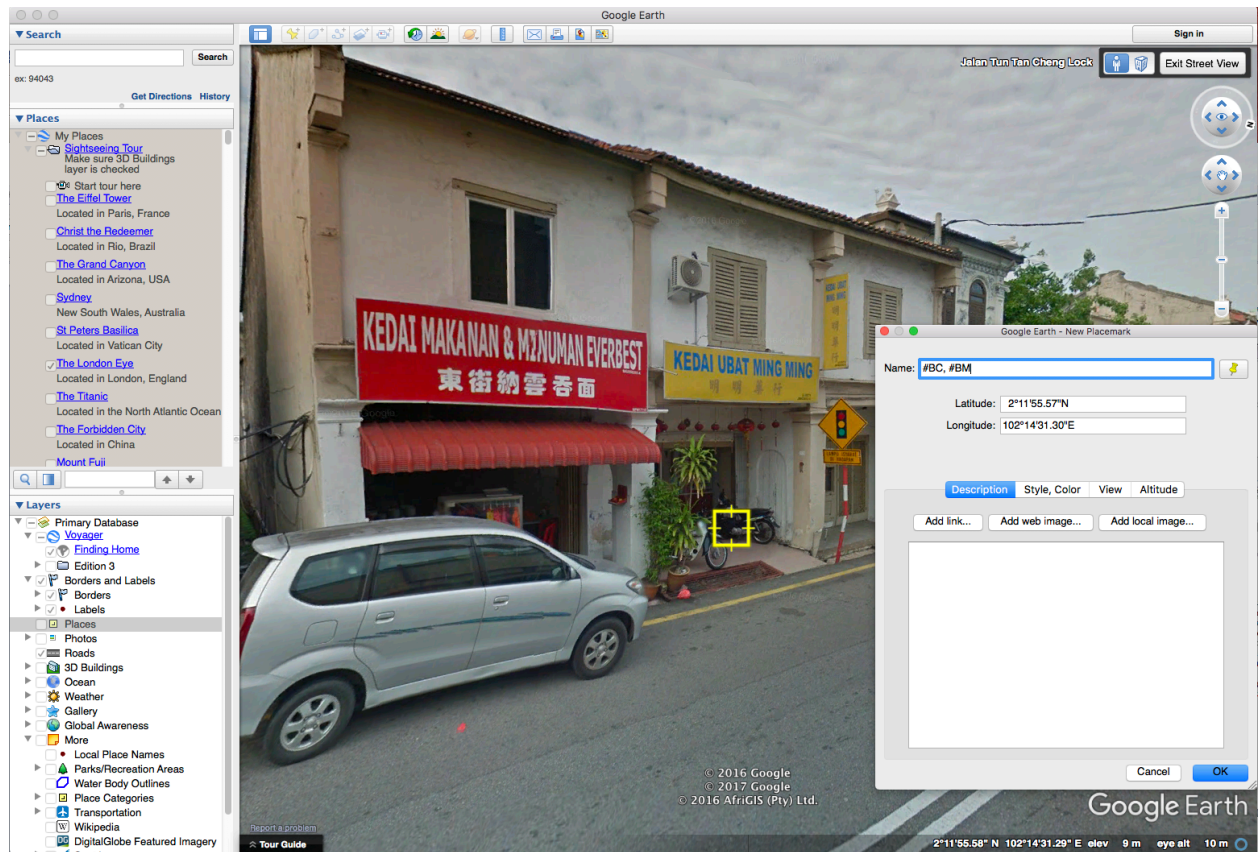
4.1 A Sample Application for Linguistic Landscape Workflow

The methods and workflow employed in this study, from sampling to visualization, have been included as a sample application for the benefit of further linguistic landscape research. Though the application does not include any suggestions for an indicator design, some sampling methods are discussed. The software used in this application includes ArcGIS 10.3 and ArcScene; these are suggestions only. The outline of this practice can be seen in the following figure:



LINGUISTIC LANDSCAPE APPLICATION WORKFLOW

4.1.1 Geo-Tag Point Set and Export KML



© DATA COLLECTION IN GOOGLE STREET VIEW

Geo-tagging a point set in Google *Street View* requires little more than a text editor into which the final exported KML can be pasted and saved. Depending on the number of tokens intended to be taken from a collection area, versioning is highly suggested to prevent data loss in the event of a crash. Frequent KML export and splicing is performed on the 4300+ geotags. Splicing of tags should be done from the open to the close of 'Placemark'. Note the location of the linguistic data housed in 'Name' in the following token:

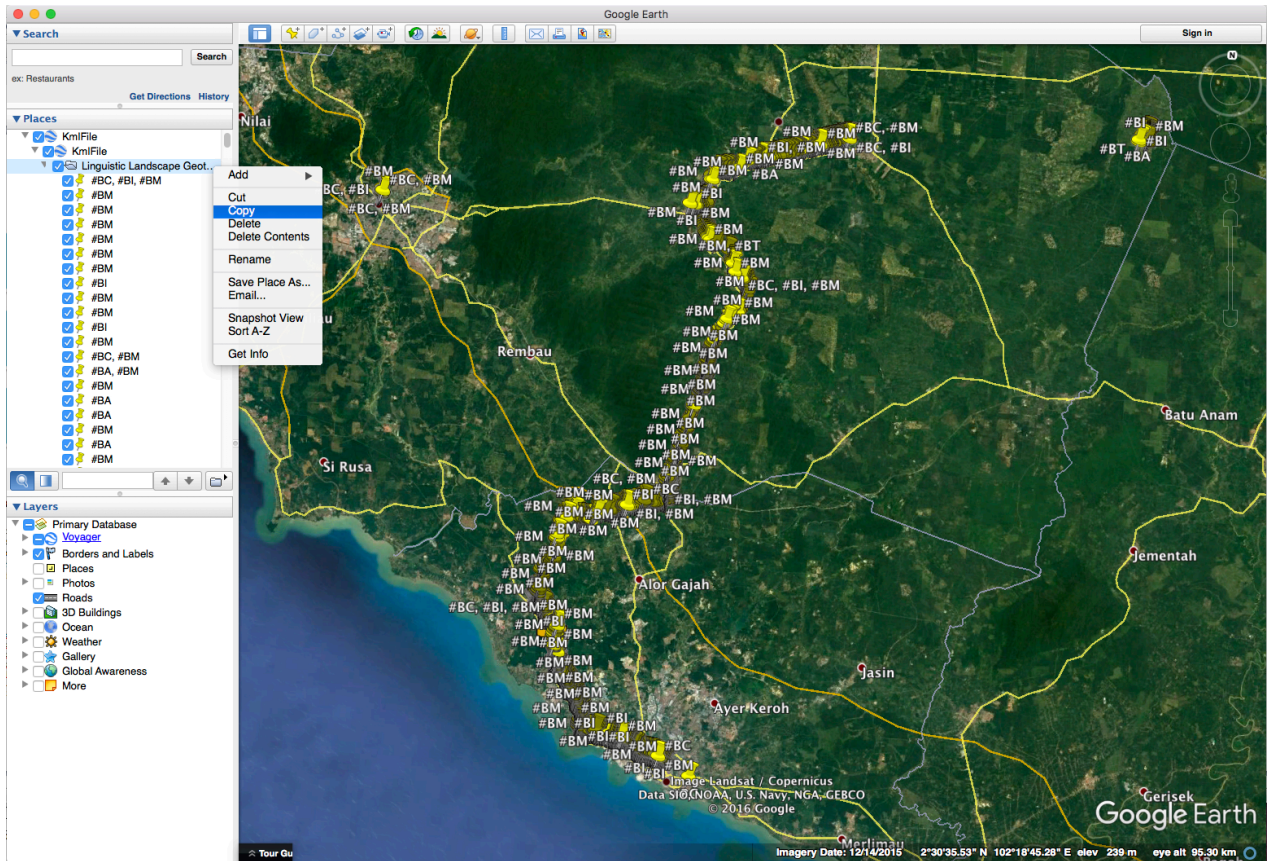
```
<Placemark>
  <name>#BC, #BM</name>
  <LookAt>
    <gx:ViewerOptions>
      <gx:option enabled="0" name="historicalimagery"></gx:option>
      <gx:option enabled="0" name="sunlight"></gx:option>
      <gx:option name="Street View"></gx:option>
    </gx:ViewerOptions>
    <longitude>102.1297264871086</longitude>
    <latitude>2.397253188567143</latitude>
    <altitude>0</altitude>
```

```

<heading>82.81599777228897</heading>
<tilt>77.66570690105503</tilt>
<range>15.6621074856617</range>
<gx:altitudeMode>relativeToSeaFloor</gx:altitudeMode>
</LookAt>
<styleUrl>#m_ylw-pushpin</styleUrl>
<Point>
  <coordinates>102.1297264871086,2.397253188567143,0</coordinates>
</Point>
</Placemark>

```

In versioning the KML geotags, it is a good practice to apply a naming convention that incorporates a date in order to time-stamp a collection, even though *Street View* coverage is only periodically updated. *Street View* coverage updates are announced for some countries and are posted by Google: <https://www.google.com/intl/en-US/StreetView/understand/> geo-tags (KML) from sample area can be seen below:

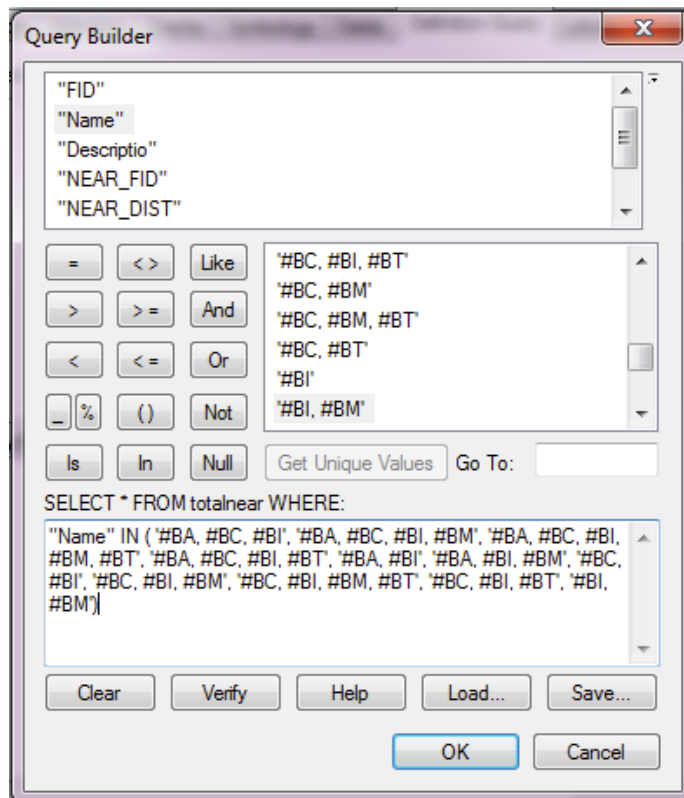


© GOOGLE EARTH KML EXPORT

4.1.2 Point Import, Point Decomposition

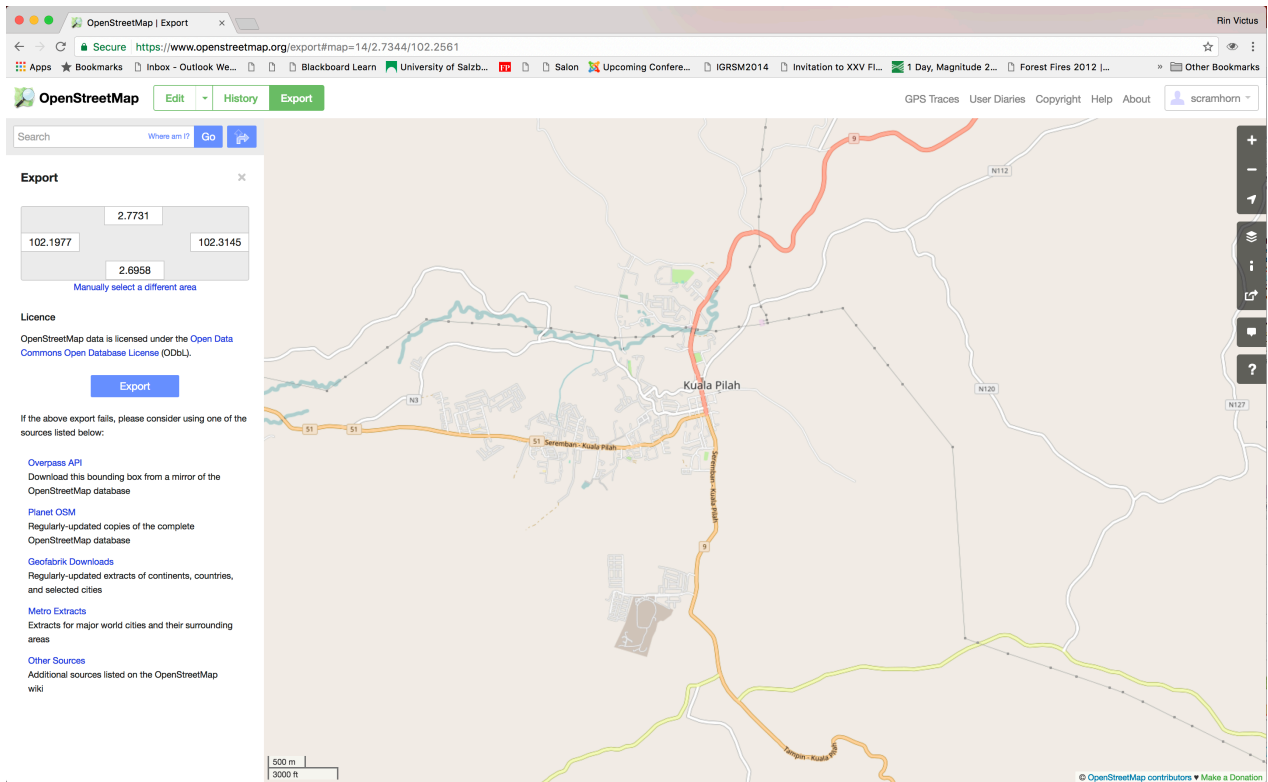
KML point sets are imported into QGIS or ArcGIS with the default spatial reference system for Google Earth is ESPG:4326, (WGS84). In either ArcGIS or QGIS, QueryBuilder is used to make the key point selections in this study for a custom metric called *lingua franca preference*. In this instance, the following SQL command calling non-monolingual instances

of English in the linguistic landscape. This query selection is exported and saved as a shapefile:



4.1.3 OSM Map Export

OSM provides the base map used for the construction of the polyline network. According to the SANET manual, a polyline network must be a continuous network “connected to each other in SANET, all the isolated polylines have to be eliminated” (Okabe, 2002a:3) This requirement imposes stringent demands in order to satisfy the constraints of a continuous graph. These demands are too optimistic for OSM Map Export. Though the completeness and accuracy for some tested areas of volunteered OSM data are comparable in quality to professional grade cartographic output (Haklay, 2008), the major problem observed in OSM data is the inconsistency of node placements at intersections, as well as the presence of ‘zero’ nodes. The Open Street Map data export limits the size of the bounding box, though larger areas can be selected by using an OSM API such as Overpass or Planet OSM. (OpenStreetMap, 2017) A base map is exported for each individual local area by bounding box export mainly for reference and for visualization purposes.



4.1.4 Network Building

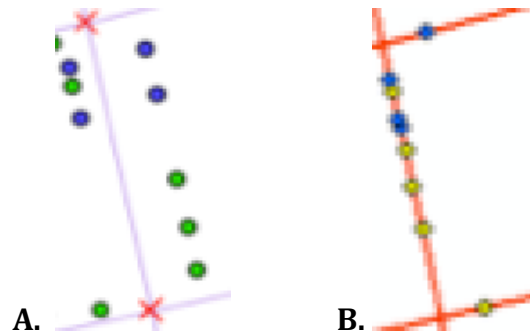
OSM map exports may be used to build the polyline network but require extensive pre-processing to clean the nodes and polylines in order to make a continuous graph. According to the SANET manual, a continuous graph “[has] to be free of intermediate or pseudo points.” (Okabe, 2009:2) The SANET manual provides instructions on this process and tools to clean OSM map data, as well as a helpful feature to find breaks in the network. One such feature produces shapefiles indicating missing or unusable network links.

In this sample application, a simplified graph is made from scratch in order to satisfy the requirements of a continuous graph but to enforce a constraint on the calculations.

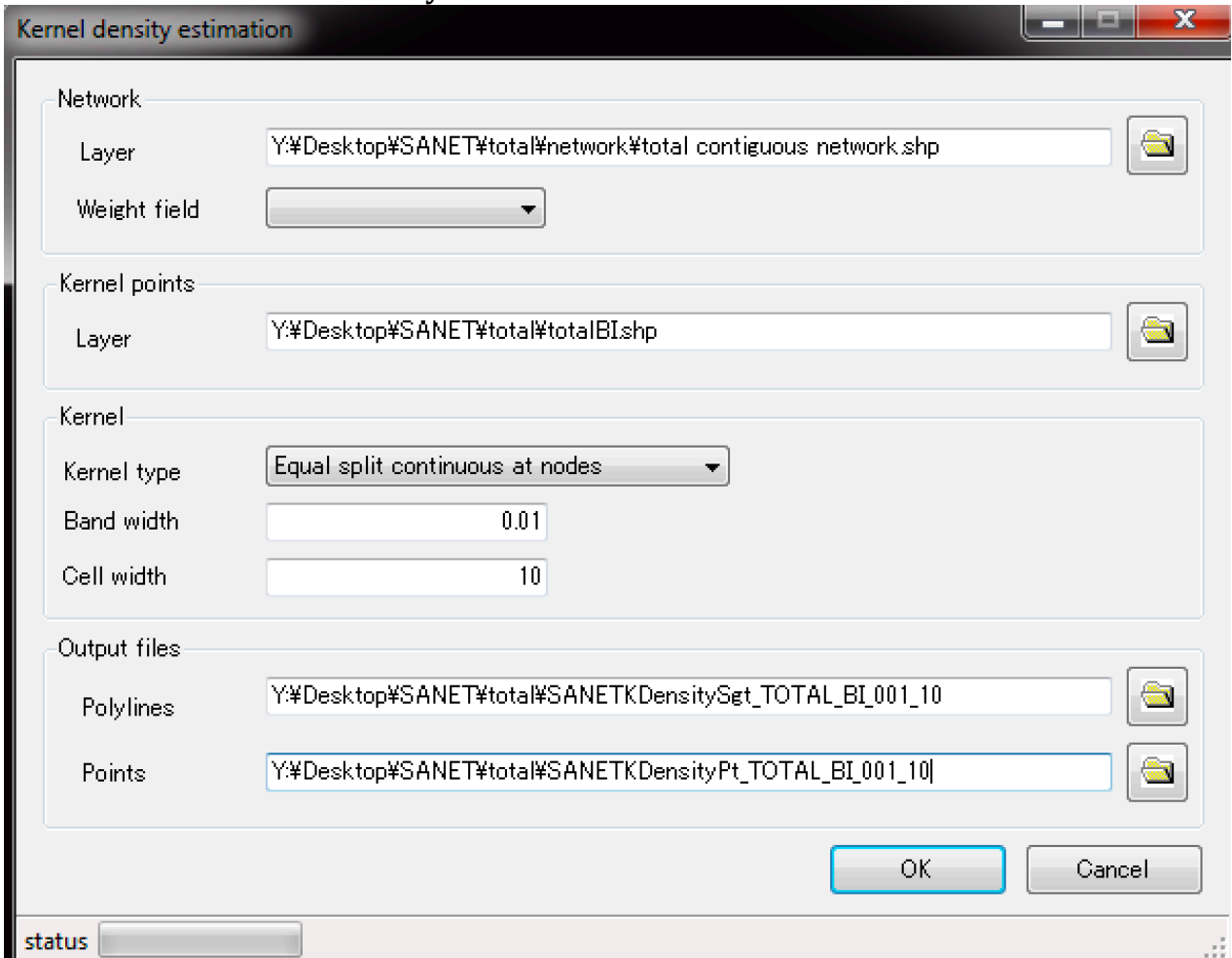
4.1.5 Map Matching

An important issue in applying the network functions to an ‘along-network’ point set is determining m for a linguistic token and to which feature an along-network token will match to the network. Along-network tokens must be assigned a network location that most appropriately matches the network with a map matching tolerance that prevents the points from adhering from incorrect locations, namely the wrong streets. Because an along-network LL dataset is near the network, but not exactly on it, the point set must be matched to the polyline network. In doing so, the offset may cause distortion to the real values of the location on the network at small scales, especially when matching to a tightly bound street network. The *Street View* sampling creates an along-network offset that may cause measuring errors resulting in measures where offset linguistic tokens near nodes tend to be

closer to perpendicular streets than their actual street of residence.



4.1.6 SANET: Kernel Density Estimation



Kernel density estimation calculations are performed at 10m , 100m, and 500m on points

along a 110-km non-intersected network using SANET Standalone 4.1. The 'Equal Split Continuous at Nodes' kernel type is most appropriate for this kind of network because of the absence of intersections; a justification for this decision was made in 3.5 SANET.

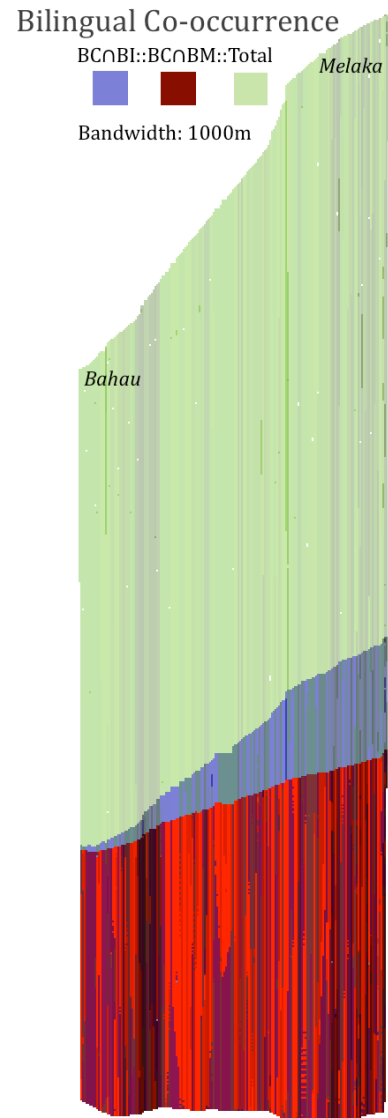
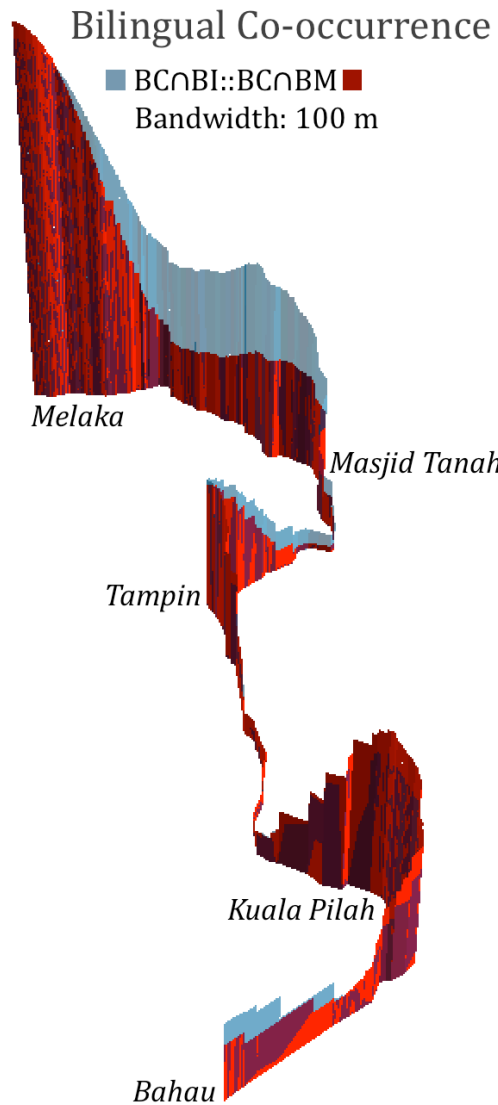
The calculation is performed at each bandwidth and with the point set of each *lingua franca*. (See Query 4.1.2) It is a best practice to save output file names indicating the parameters and linguistic sets used for quick reference.

4.1.7 ArcScene: Base Heights Adjustment and Extrusion

The results of a SANET kernel density estimation calculation are given in 3-D line graphs. While ArcScene is used in this thesis to visualize these 3-D line graphs, the instructions included for adjusting and filling the graphs included in the SANET manual employ ArcScene. The first suggestion indicates 'Extrusion', which colorizes the vertical face of the graph. Depending on the bandwidth, the local intensity may be too high to fit with both the base OSM map and the peaks of the kernel density graph itself into a single scene. In order to reduce this intensity and create a presentable scene, an acceptable scale of z-elevation reduction is by employing an expression that reduces the overall height of the graph substantially. The use of such an expression can be seen in ArcScene where the mean "[Average]*0.0005". This will reduce the intensity but retain the local trend variation. However, these base heights adjustments depend on the local intensity measures and the bandwidth. The SANET Manual provides additional support for visualizing 3-D KDE results.

4.1.8 ArcScene: Overlay and Transparency

In order to visualize co-location patterns with KDE measures, these graphs must be modified to serve the purposes of this thesis. Visualizing trends as stacked line graphs requires first that identical calculations be performed on two different point sets. This means identical bandwidth, identical cell size using the same network. Once both measures have been performed, overlain, extruded and adjusted for base heights (4.1.7), adjusting the transparency for the dominant measure in the trend allows for a clear stacked line graph to present co-location trends. This adjustment may require choosing a light color and choosing a dark color for the non-dominant trend. There is, of course, a problem visualizing the points where the when non-dominant density measure becomes submerged. Co-location KDE measures may be employed as cartograms because the base maps or additional cartographic elements are not included to prevent occlusion or when dramatic trend graph adjustments are made. Please note the following examples:



A. B.

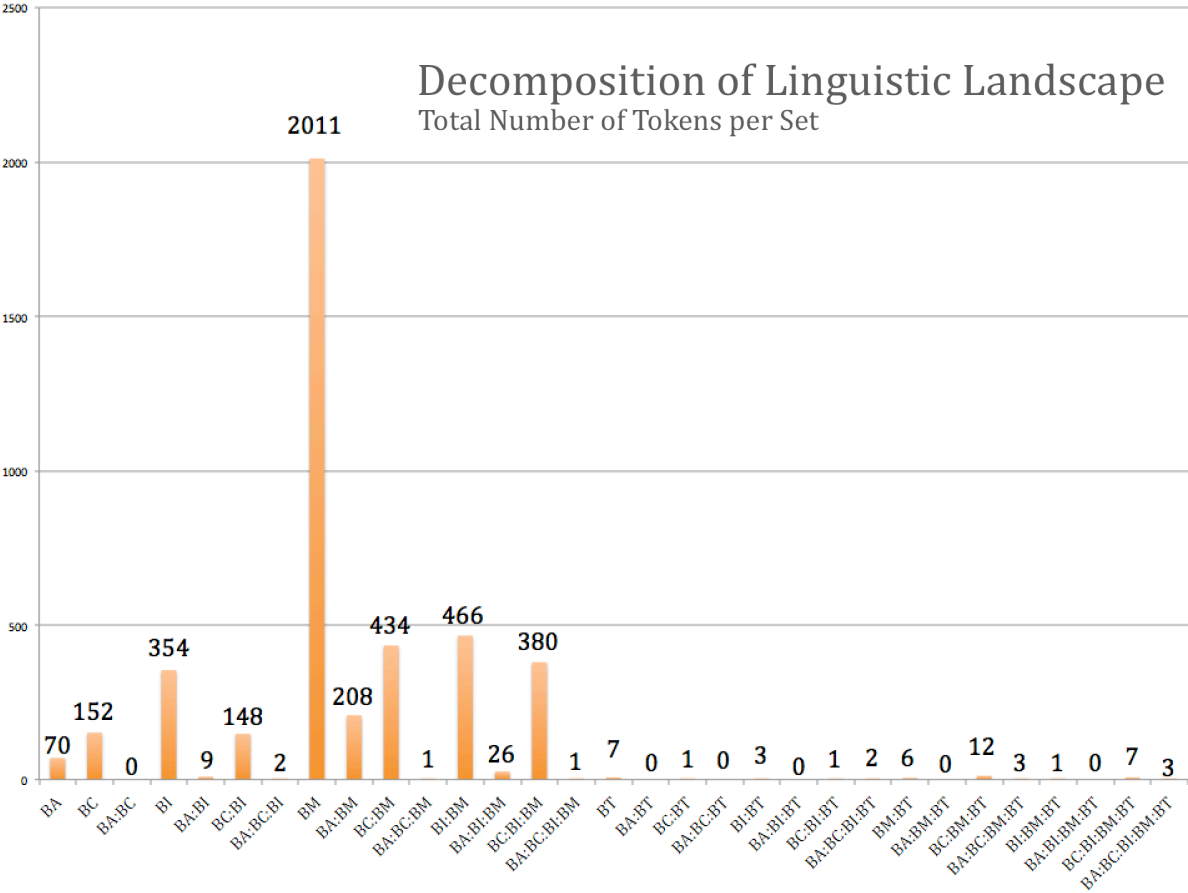
RESULTS

5.1 Sample application results

In this section, the results for this study are visualized both as non-spatial aggregate metrics as well as local and regional trend maps. The aggregate statistics are shown for linguistic landscape token total, instance *Total*, *Monolingual Composition*, *Co-occurrence*, *Affiliation*, and Greenberg's Linguistic Diversity. The following results are yielded in both aggregate graphical form and indicate results taken from a number of sample areas. Local and regional trend maps are given for Kuala Pilah, Tampin, Melaka, and a greater regional study area, all visualized as trend maps. Geographical coordinates are displayed for centers of local intensity. Included in the trend map is bandwidth, network edge length, color ramp, *OpenStreetMap* base map, a description of the *lingua franca* set (intersecting or non-

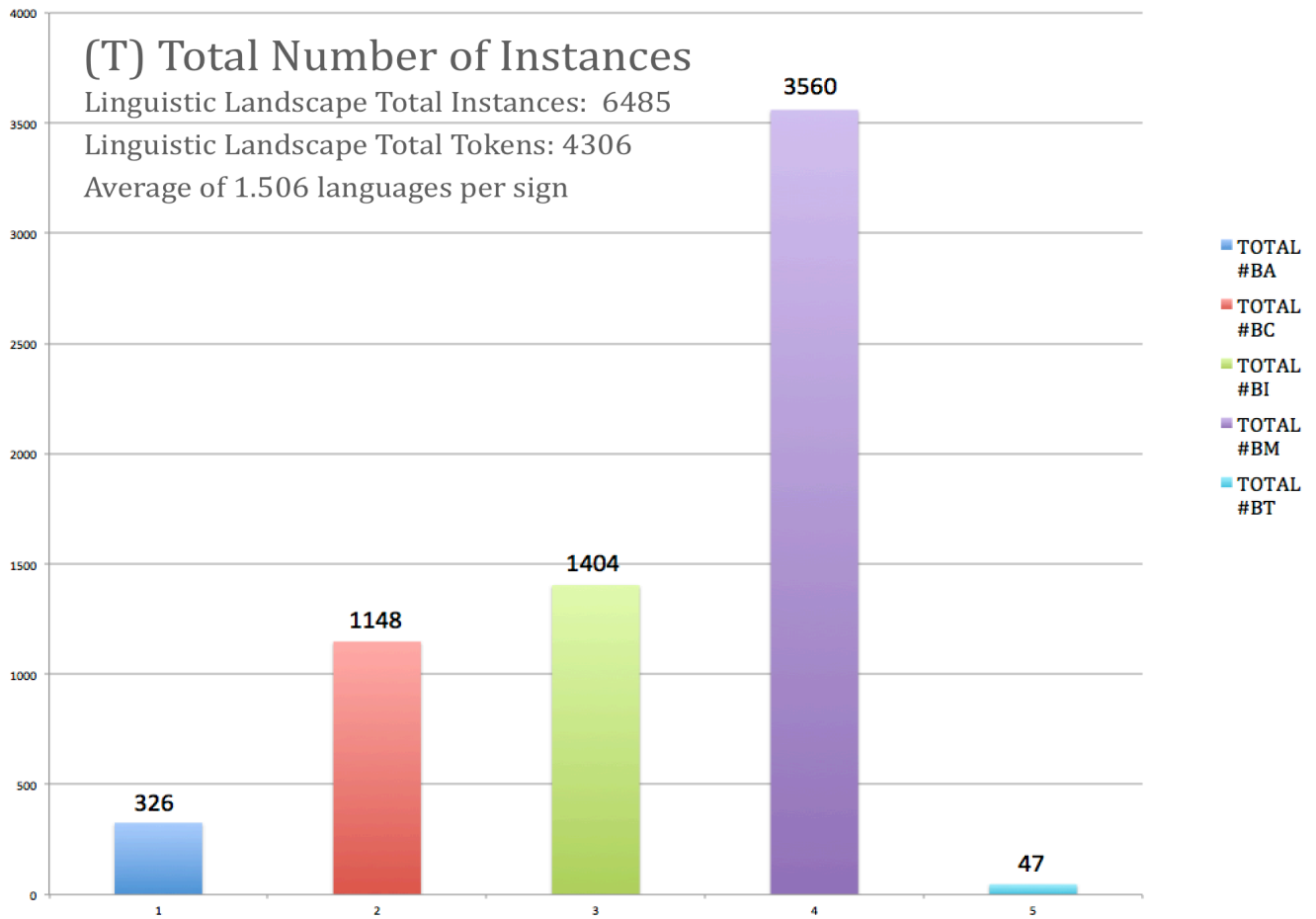
intersecting), and kernel density measures centered at local intensity points. To provide complete results for all spatial and non-spatial metrics in each location is beyond the scope of this study.

5.1.1 Total tokens: Decomposition of the linguistic landscape



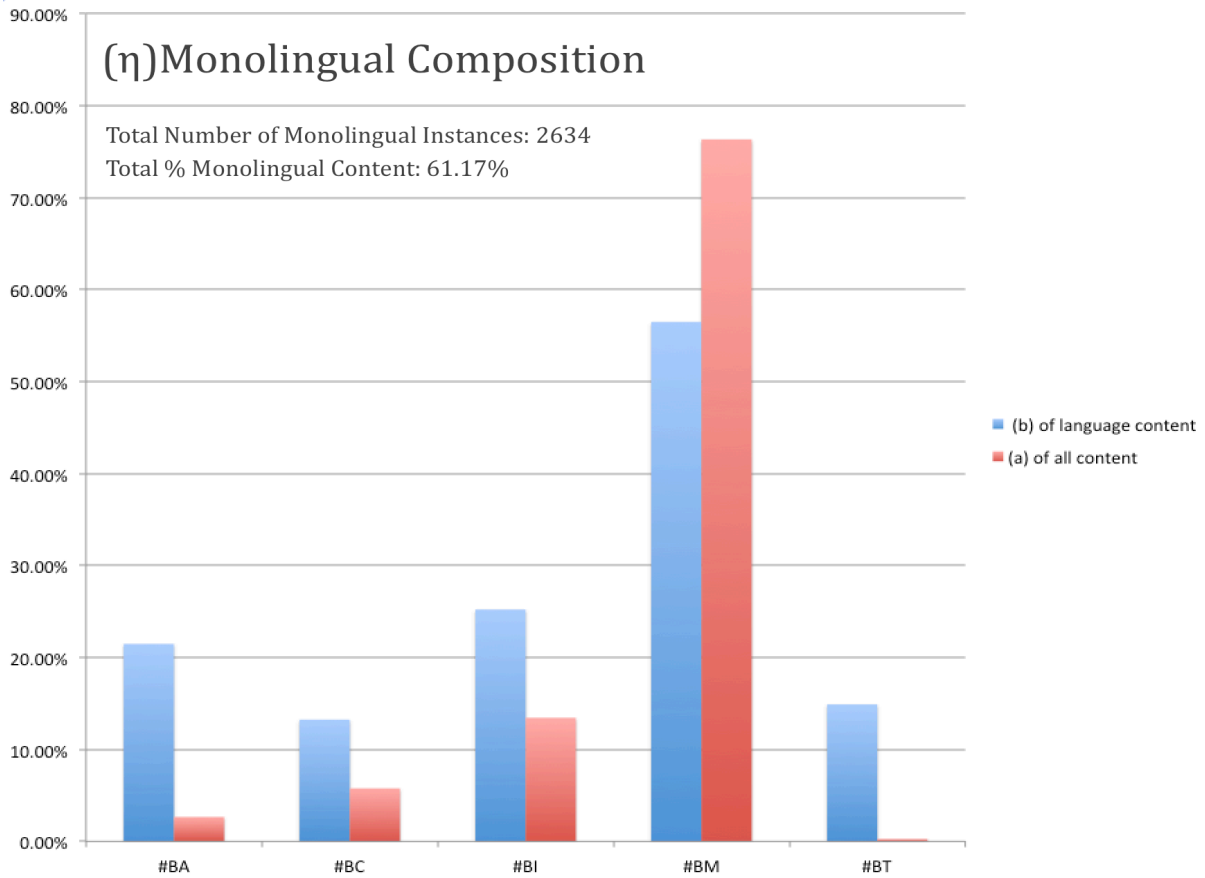
The vast majority of Malaysia’s linguistic landscape is monolingual Malay at 46.70%. This seems to be in line with the status of Malay as the top-down state-sanctioned official language as well as representative of the majority population in peninsular Malaysia. Additionally, BC:BM (10.08%), BI:BM (10.82%), BC:BI:BM (8.82%), and BI (8.22%) constitute significant portions of the linguistic landscape. Including the groups BA:BM (4.83%), BC (3.53%), BC:BI (3.53%), and BA (1.63%), these groups total 98.16% of the total sample. BT was not included in any group constituting the bulk of the linguistic landscape.

5.1.2 Total instances



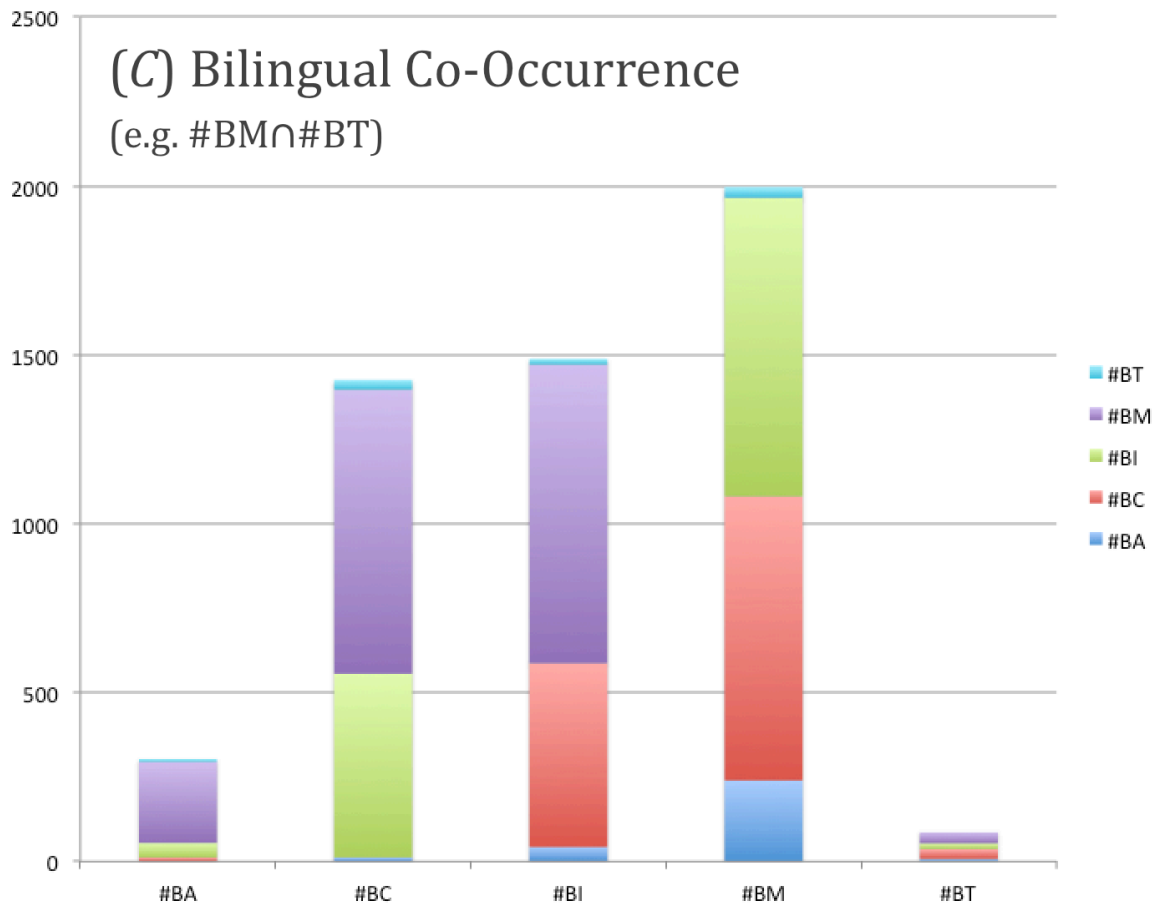
In (T), the instance totals are decomposed by language (script). Of the 4306 tokens, 6485 linguistic instances were recorded for a total token average of 1.506 languages per sign. BT is the least represented script in the linguistic landscape, and seems to be underrepresented in the linguistic landscape.

5.1.3 Monolingual composition



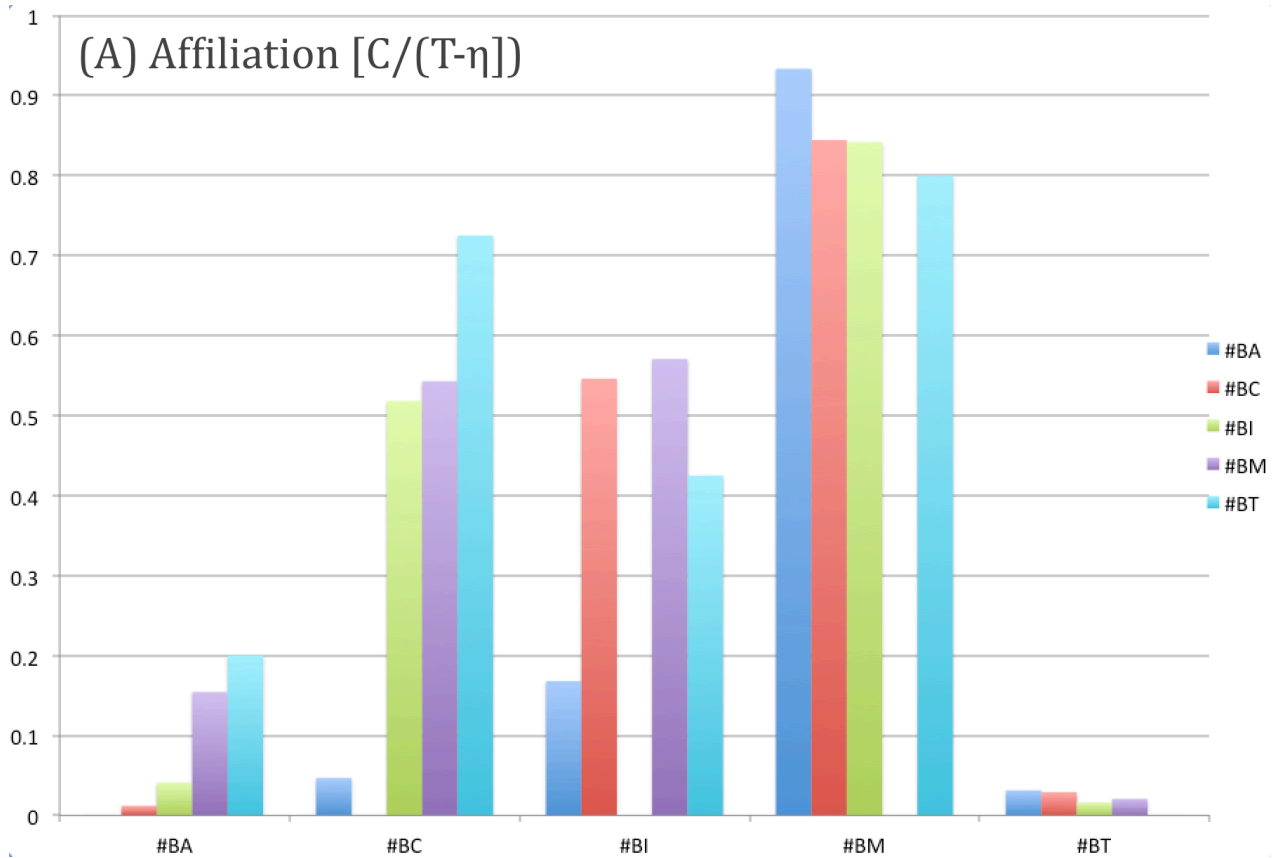
In this metric, the total monolingual content of the linguistic landscape sample was measured at 61.17% with 2634 monolingual tokens recorded. Malay counts for 76.34% of the monolingual content of the study area and 46.70% of all total study area content in the study area. This metric shows great variation in local intensity and varies wildly depending on the sample area, with a much higher multilingual content and a lower monolingual Malay quotient for urban areas.

5.1.4 Bilingual Co-occurrence



The (C) Co-occurrence metric is a bar graph adapted from a matrix of linguistic unions. One initial observation is the high occurrence of unions made by #BA with #BM, seen in Column 1. The sample area indicates that that multilingualism occurs most often with the #BC, #BI, and #BM scripts. #BA occurs almost always with #BM, occasionally with #BI, but very rarely with #BC.

5.1.5 Affiliation



The results for the *Affiliation* measure have successfully isolated the level of attraction between each language configuration in the Malaysian linguistic landscape. It can most notably be seen that Tamil (BT) shows a greater level of attraction to Chinese and Malay and seems to be less affiliated with #BI. This may be due to #BT occurring in zones where Malay is the *lingua franca* and rarely within the urban study regions of Melaka. Additionally, the use of English as a language of status may be interpreted from the affiliation of BI with BA. The most notable level of affiliation is that of the inclusive Jawi script with Malay.

5.1.6 Greenberg's Linguistic Diversity Index

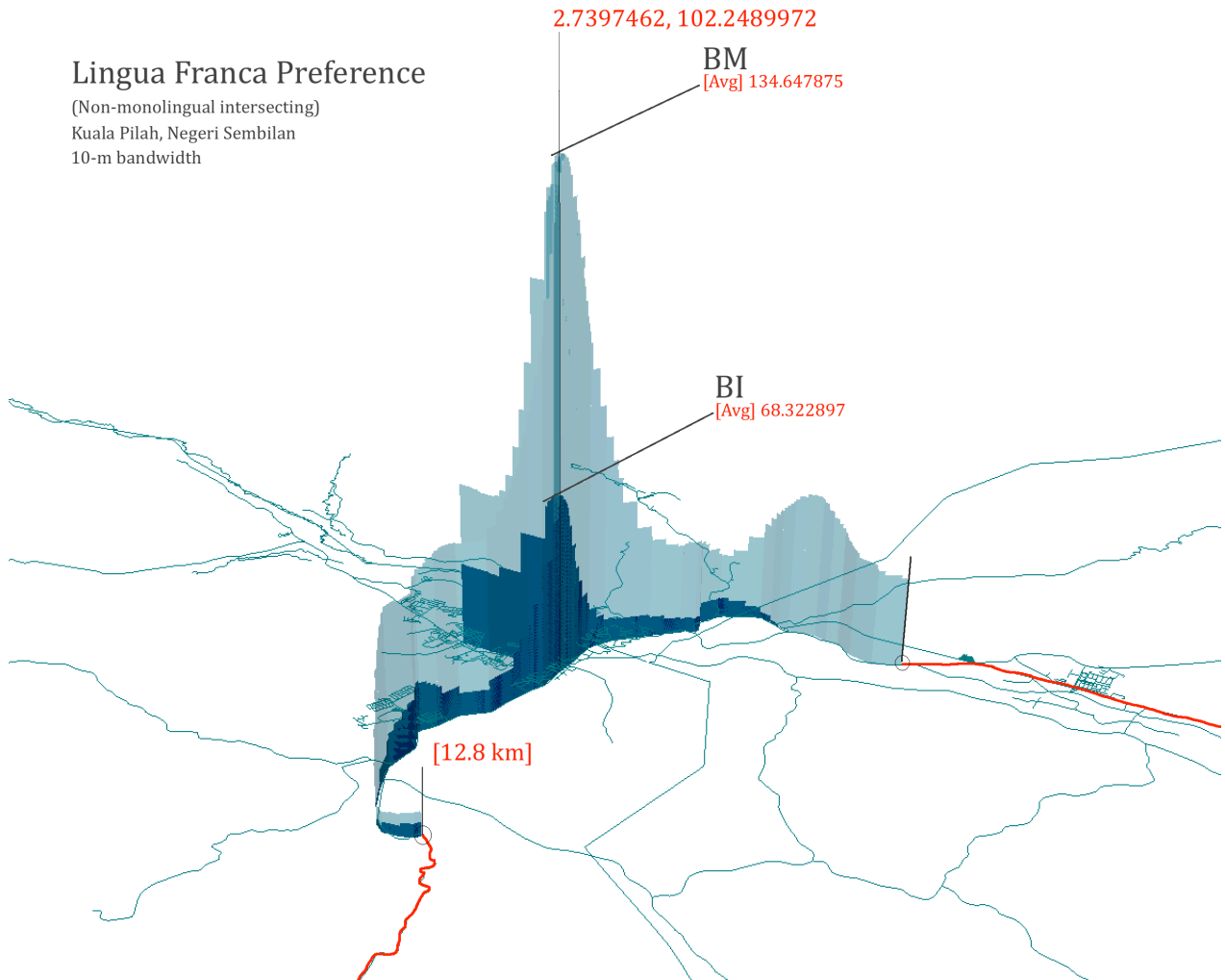
Greenberg's Linguistic Diversity Index

$$A = 1 - \sum_i (i^2) = 0.740$$

The Linguistic Diversity Index was calculated by measuring the aggregate linguistic landscape tokens by using the *Split-Personality Method* of Greenberg's Linguistic Diversity Index (Greenberg, 1955). This method "count[s] every speaker of two languages as two people, every trilingual as three, and so on." (Greenberg, 1955:111) In applying this method to the linguistic landscape, each class is counted separately, both monolingual and polylingual configurations. The calculation resulted in a Linguistic Diversity Index of 0.740. The UNESCO World Report indicates an index of .758 (UNESCO, 2009).

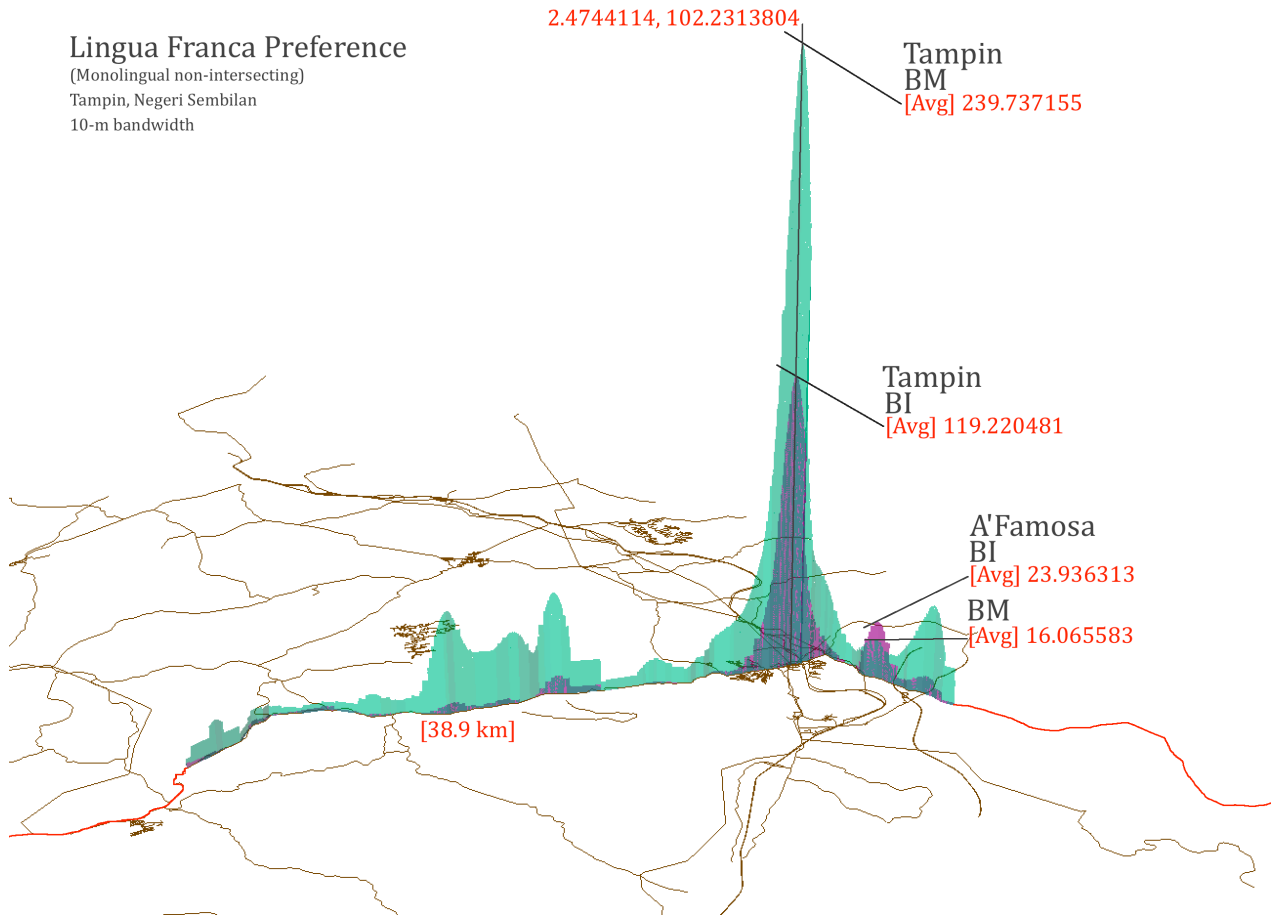
It should be noted that the Malaysian linguistic landscape has virtually no content representative of Orang Asli languages. If it were to represent the Orang Asli languages, may result in a slightly higher index given then small number of speakers in that group. It may be inferred from this result that the linguistic landscape population counts may be a viable linguistic diversity indicator.

5.2 Kuala Pilah



In this local trend map, the results for Kuala Pilah, a town of 18,000 located in Negeri Sembilan, clearly exhibit a local intensity in the urban center. The 'Equal Split Continuous' kernel function was used to estimate density of non-monolingual intersecting lingua franca co-locations. A 10-m bandwidth is used on the two sets of a 12.8-km long corridor. Non-monolingual eliminates both monolingual #BI and #BM from the sets and counts the unions [BI:BM:~] for both BI and BM. The results indicate a presence of but not preference of English in the urban center with Malay *lingua franca* dominant outside the urban area where the urban area tapers off.

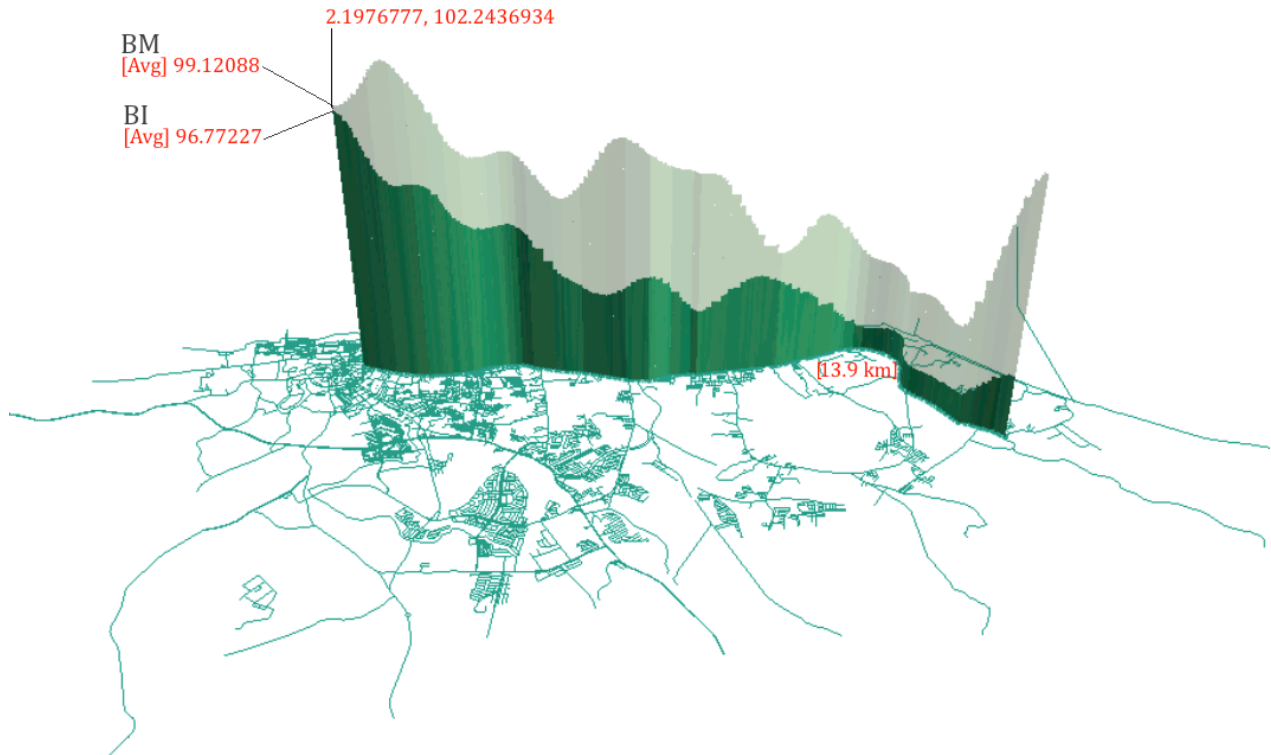
5.3 Tampin



In this local trend map, the results for Tampin, a town of 58,000 located in Negeri Sembilan, clearly exhibit a local intensity in the urban center. The 'Equal Split Continuous' kernel function was used to estimate density of monolingual intersecting lingua franca co-locations. A 10-m bandwidth is used on the two sets of a 39.8-km long corridor. Monolingual *Preference* includes both monolingual #BI and #BM from the sets and counts the unions [BI:BM:~] separately for both BI and BM; mutual unions are excluded. The results indicate a presence of English in the urban center with the Malay *lingua franca* dominance tapering off outside the urban area. Additionally, there is a local intensity of #BI outside the town center designated 'A'Famosa'. This is an area of resorts known for international tourism, an area well-advertised with monolingual #BI tokens.

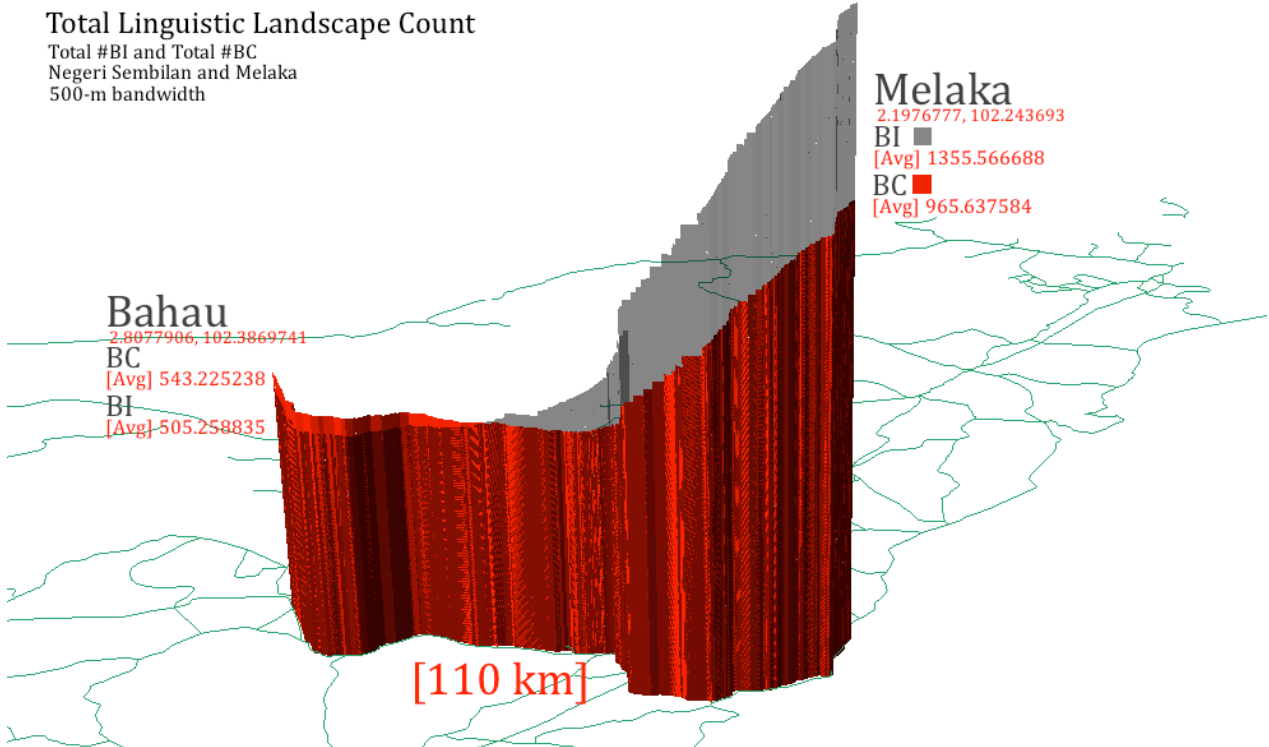
5.4 Melaka

Lingua Franca Preference
(non-monolingual intersecting)
Melaka, Melaka State
10-m bandwidth



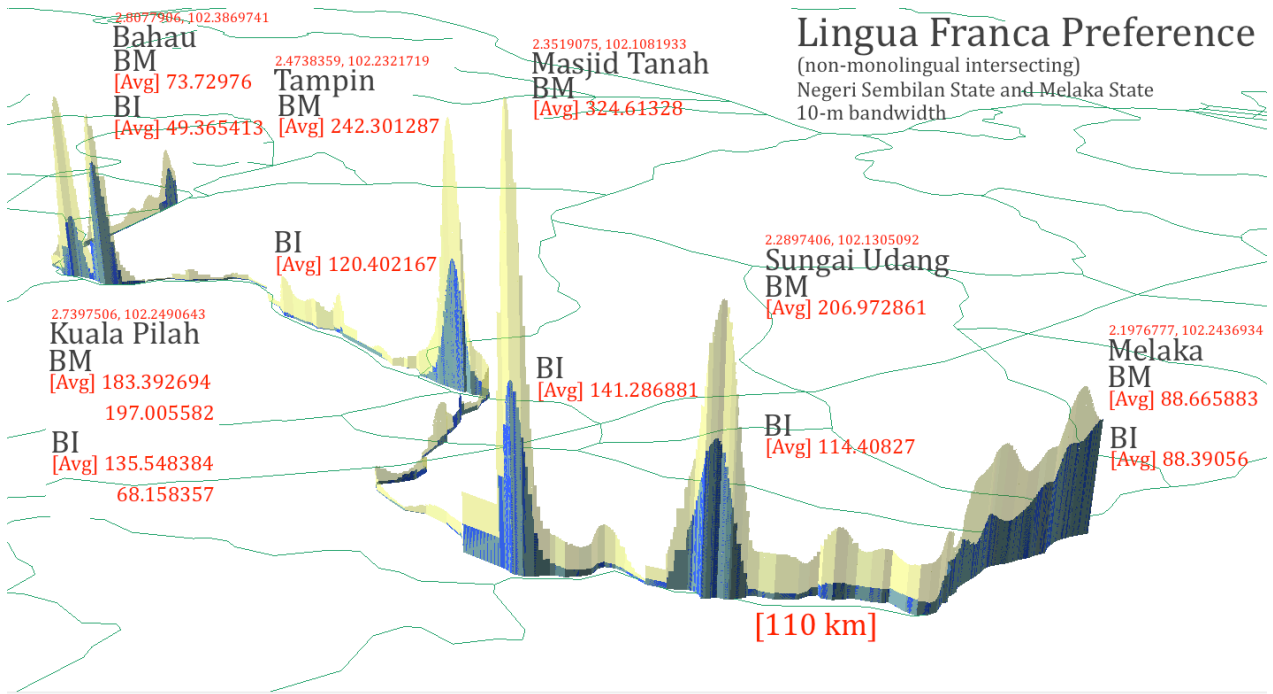
In this local trend map, the results for Melaka, a city of 484,000 located in Melaka State, do not clearly exhibit a local intensity in the urban center. The 'Equal Split Continuous' kernel function was used to estimate density of non-monolingual intersecting *lingua franca* collocations. A 10-m bandwidth is used on the two sets of a 13.9-km long coastal corridor. Non-monolingual eliminates both monolingual #BI and #BM from the sets and counts the unions [BI:BM:~] for both BI and BM. Instead of a single point of local intensity, such as the local intensity spikes seen in Tampin and Kuala Pilah, there is a gradual increase in density for both #BI and #BM tokens, culminating in a point of equal-use in the city center.

5.5 Composite



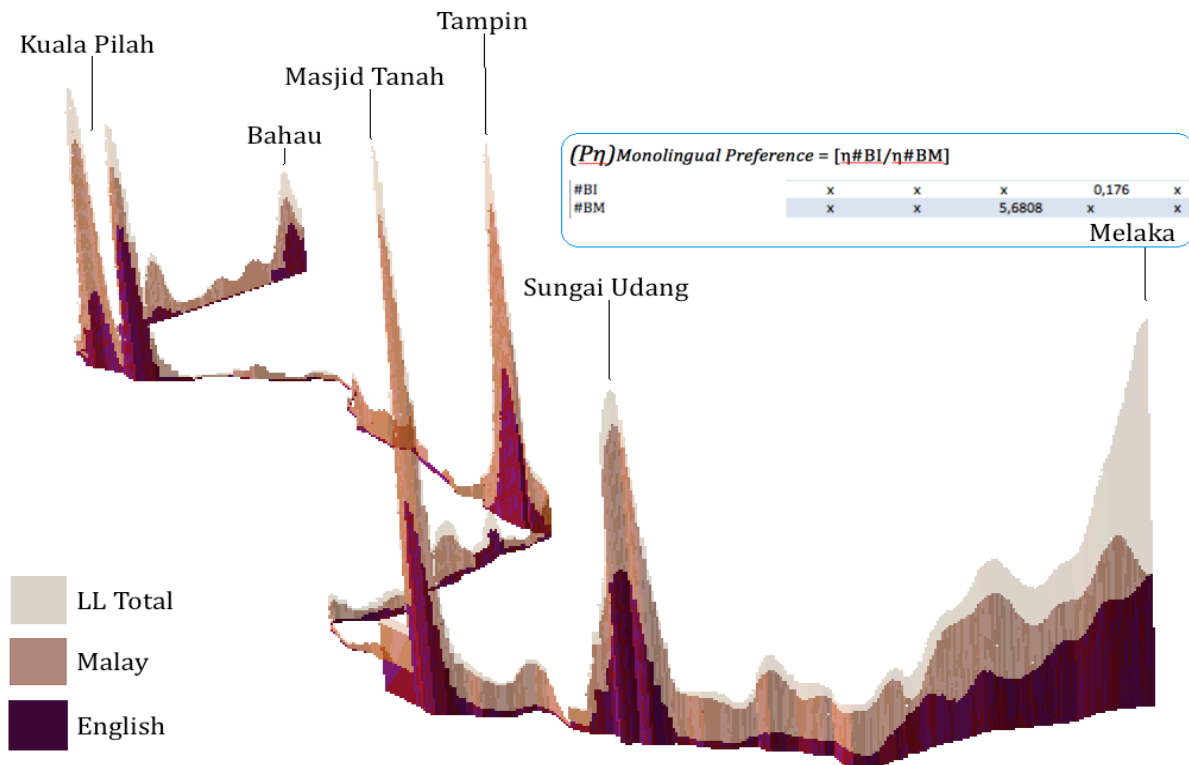
In this composite trend map, the 'Equal Split Continuous' kernel function was used to estimate density of co-location patterns for the (T) Total #BI and #BC counts. A 500-m bandwidth is used on the two sets of a 110-km corridor. A larger bandwidth was used in order to smooth intensities and show broader trends. The gradual increase in linguistic landscape density can be noted. A slight majority #BC can be seen in the easterly extent of Bahau while a much greater #BI majority content be noted in the westerly extent of Melaka. A directional trend can be noted from Bahau to Melaka in which one can note the increase of English over Chinese presence.

It may be noted that the visualization may be oblique at the easterly extent and westerly extent due to the shape of the sample area. Such obliqueness was accounted for and compensated for with transparency changes. This trend map was adjusted to best fit the trend graph with an acceptable view of the OSM base map. When projecting a trend, finding the right degree of z-line reduction and angle of view may require much manipulation and adjustment.



In this composite trend map, the 'Equal Split Continuous' kernel function was used to estimate density of co-location patterns for non-monolingual intersecting *lingua franca* co-locations. A 10-m bandwidth is used on the two sets of a 110-km long corridor. Kernel density estimations are made for Bahau, Kuala Pilah, Tampin, Masjid Tanah, Sungai Udang, and Melaka.

It may be noted that the visualization may be oblique at the easterly extent and westerly extent due to the shape of the sample area. Such obliqueness was allowed and compensated with transparency changes and arithmetic reduction of kernel density measure lines. However, it seems more appropriate to reduce small bandwidth trend lines further due to the size of local intensity spikes.



In a final sample application trend map, the total 110-km sample area can be seen with triple co-location selection of linguistic landscape *Preference* sets. The selection in this trend map is monolingual Malay (#BM), monolingual English (#B1), and total linguistic landscape count—mono- and multilingual). With the option to select the quantitative linguistic landscape metric, the user's color ramp indicates the script selection and the drop-down window gives local statistics for the metric. The user can select lengths of corridor and multiple corridors, select bandwidth, cell size, a base map, transparency, and network corridor detail.

C O N C L U S I O N S

6.1 Summary

The use of SANET network kernel density estimation method to analyze trends in the linguistic landscape has yielded intriguing insights into the distribution of linguistic elements in Malaysia's landscape. Additionally, the use of Google *Street View* has proved to be an extremely effective tool for the data collection of Malaysian linguistic landscape tokens.

SANET kernel density functions have been able to show directional trends and changes in *Lingua Franca Preference* co-locations of these trends. Directional trends were noted from Bahau to Melaka in which one can note the increase of English over Chinese presence using the 500-m bandwidth. It was noted that Melaka town has nearly an equal use of English as Malay as a lingua franca at the western extent of the study area. Additionally, local intensities were able to detect the English-preferring enclaves of A'Famosa, near Tampin. Aggregate measures of Linguistic Diversity Index have shown to be 0.740, UNESCO World Report calculation for Malaysia at 0.758. ^{1e}

6.2 Constraints

Google *Street View* offers a vast potential for linguistic landscape data collection. However, at present, there are limitations in the employ of total sampling method. Currently, this process is a supervised process where each token is keyed by hand. Ultimately, the automation of the data collection process is required for the vast sampling required by this method. Should a full application be modeled on this method, the number of tokens may be astronomical if it were to cover all regions of the country, well beyond the feasibility of human labor. Below is a conservative estimate based on the total road network of Malaysia:

Number of Tokens in Study Area: 4306

Size of Study Area: 110 km

Average Number of Tokens per km: 39.14545

¹Size of Road Network of Malaysia: 144,403 km

Estimate of Total Tokens: 5652721 tokens

The solution to such an overwhelming data collection task would be to utilize an unsupervised training method, specifically with an optical character recognition software (OCR). Though current OCR algorithms are capable of character recognition with up to 99% accuracy on some scripts (Holly, 2009), there are still obstacles to training 'in the wild'. OCR technology is ideally suited for "[quickly] making textual versions of printed documents, e.g. book scanning for Project Gutenberg, making electronic images of printed documents searchable, e.g. Google Books, and defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR." (Resig, 2009) Google OCR is based on an open-source algorithm called Tesseract. (Smith, 2007) Training an algorithm to select individual linguistic

tokens in *Street View* coverage would encounter multiple problems; these same problems are employed by CAPTCHA systems to hinder OCR probability of detection, namely: skewness, speckling, binarisation, lines, zoning, line-word detection, script detection, character isolation, and aspect ratio. (Sezgin et al, 2004), (Gupta et al, 2007), (Trier et al, 1995), (Milyaev et al, 2013), (Pati et al, 1987), (Smith, 2007). Of these obstacles, ‘script detection’ seems to present a substantial training challenge to evaluating multilingualism and linguistic diversity measures. This ‘script detection’ problem specifically refers to detecting languages in multilingual images where the use of multiple OCR requires zonal selections. In trial attempts of using Google *Docs* OCR software to test detection capabilities, a number of tokens/signs were tested:



𑖀𑖁. .



𑖀𑖀𑖀𑖀: – DAUGIAU



POLIILINIK HIDAYAH

A போளிகிEக் ஹிடாயா



ADZ UAN كEDATTAYAR DANSERVISKENDERAAN ʋлах
TYR s

WORLD CLASS MOTOROL

Fe(ON-li "N366 Göre " .

In addition to dealing with detection problems, unsupervised collection of *Street View* imagery requires a number of advances in identifying the borders of a single linguistic token. As defined by Cenoz and Gorter, "any piece of text within a spatially definable frame" constitutes a single instance, from "small handwritten stickers to huge commercial billboards." (Gorter, 2006)

6.3 Further research

While it is generally agreed that there is no shortage of interest in studies of the linguistic landscape (Gorter, 2013), the development of new methods of analysis and approaches such as the one in this study will hopefully increase interest of quantitative analysis of linguistic landscape. Further quantitative approaches and the analysis of further linguistic landscape criteria will very likely lead to intriguing developments. One possible research development could utilize SANET kernel density estimations on metadata from geo-tagged social media.

B I B L I O G R A P H Y

- Backhaus, P. (2005). *Signs of multilingualism in Tokyo: A diachronic look at the linguistic landscape*. International Journal of the Sociology of Language, 175/176, 103–121.
- Backhaus, P. (2006). *Multilingualism in Tokyo: A look into the linguistic landscape*. International Journal of Multilingualism, 3 (1), 52–66.
- Backhaus, P. (2007). *Linguistic landscapes: A comparative study of urban multilingualism in Tokyo*. Clevedon, UK: Multilingual Matters.
- Backhaus, P. (2008). *The linguistic landscape of Tokyo*. In M. Barni & G. Extra (Eds.), Mapping linguistic diversity in multicultural contexts (pp. 311–333). Berlin, Germany: Mouton de Gruyter.
- Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. Harlow Essex, England: Longman Scientific & Technical.
- Barni, M., & Bagna, C. (2009). *A mapping technique and the linguistic landscape*. In E. Shohamy & D. Gorter (Eds.), Linguistic landscape: Expanding the scenery (pp. 126– 140). New York, NY: Routledge.
- "Basic OCR in OpenCV | Damiles". Blog.damiles.com. Retrieved 2013-06-16.
- Ben-Rafael, Eliezer; Shohamy, Elana; Hasan Amara, Muhammad; Trumper-Hech, Nira (2006) *Linguistic Landscape as Symbolic Construction of the Public Space: The Case of Israel*, International Journal of Multilingualism, 3:1, 7-30
- Boeing, G. 2016. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." Manuscript under review. doi:10.2139/ssrn.2865501
- "Breaking a Visual CAPTCHA". Cs.sfu.ca. 2002-12-10. Retrieved 2013-06-16.
- Briscoe, Ulla (2009) *Geolinguistics GIS Applications: Aspects of Data Quality in Mapping Lesser-Used Languages*. (UNIGIS) MSc Thesis: Centre for GeoInformatics (Z_GIS), Salzburg University, Austria
- Cartwright, D. (2006). *Geolinguistic analysis in language policy*. In T. Ricento (Ed.), An introduction to language policy (pp. 194–209). Malden, MA: Wiley-Blackwell.
- Cenoz, Jasone; Gorter, Durk (2008) *The linguistic landscape as an additional source of input in second language acquisition* IRAL - International Review of Applied Linguistics in Language Teaching, 2008, Vol.46 (3), pp.267-287 [Peer Reviewed Journal] Walter de Gruyter GmbH & Co. KG
- Cohen, Saul B., and Nurit Kliot. 1992. *Place-names in Israel's ideological struggle over the administered territories*. Annals of the Association of American Geographers 82 (4), 653–680.
- Curtin, Kevin M. (2007) *Network Analysis in Geographic Information Science: Review, Assessment, and Projections*. Cartography and Geographic Information Science, 34:2, 103-111, DOI: 10.1559/152304007781002163
- Dijkstra, E. W. (1959). *A note on two problems in connexion with graphs* (<http://www-m3.ma.tum.de/twiki/pub/MN0506/WebHome/dijkstra.pdf>). Numerische Mathematik 1: 269–271.
- Dixon, Philip M. (2002) *Ripley's K function*. Encyclopedia of Environmetrics Volume 3, pp 1796–1803. Edited by Abdel H. El-Shaarawi and Walter W. Piegorisch John Wiley & Sons, Ltd, Chichester

ESRI Support. *GIS Dictionary*. Accessed on January 3, 2017 <http://support.esri.com/other-resources/gis-dictionary/term/variance>

Gorter, D. (Ed.). (2006). *Linguistic landscape: A new approach to multilingualism*. Clevedon, UK: Multilingual Matters.

Gorter, D. (2009). *The linguistic landscape in Rome: Aspects of multilingualism and diversity*. In R. Bracalenti, D. Gorter, I. Catia, F. Santonico, & C. Valente (Eds.), *Roma multiethnica (I cambiamenti nel panorama linguistico/changes in the linguistic landscape)* (pp. 15–55). Rome, Italy: Edup SRL.

Gorter, D., Aiestaran, J., & Cenoz, J. (2012). *The revitalization of Basque and the linguistic landscape of Donostia-San Sebastián*. In D. Gorter, H. F. Marten, & L. Van Mensel (Eds.), *Minority languages in the linguistic landscape* (pp. 148–163). Basingstoke, UK: Palgrave-Macmillan.

Gorter, D., & Cenoz, J. (2007). *Knowledge about language and linguistic landscape*. In J. Cenoz & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Knowledge about language* (2nd ed., vol. 6, pp. 343–355). New York, NY: Springer

Gorter, D., Marten, H. F., & Van Mensel, L. (Eds.). (2012). *Minority languages in the linguistic landscape*. Basingstoke, UK: Palgrave-Macmillan.

Graham, Mark; Zook, Matthew (2013) *Augmented realities and uneven geographies: exploring the geolinguistic contours of the Web*. *Environment and Planning A*, Vol. 45, pp. 77-99.

Gupta, Maya R.; Jacobson, Nathaniel P.; Garcia, Eric K. (2007). "OCR binarisation and image pre-processing for searching historical documents."(PDF). *Pattern Recognition*. 40 (2): 389.

Haklay, M. (2008) *How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets* *Environment and Planning B: Planning and Design* Volume 37, pp. 682-70

Holley, Rose (April 2009). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". *D-Lib Magazine*. Retrieved 5 January 2014.

"How To Crack Captchas". *andrewt.net*. 2006-06-28. Retrieved 2013-06-16

Hung, Helen Ting Mu (2013) *Language, Identity and Mobility: Perspective of Malaysian Chinese Youth* *Malaysian Journal of Chinese Studies*, 2013, 2(1): 83-102

Inoue, Fumio (2012) *Improvements in the sociolinguistic status of dialects as observed through linguistic landscapes—Utilization of Google Maps and Google Insights* *Dialectogia* Issue 8 (2012), 85-132.

Kasanga, Luanga Adrien (2012) *Mapping the linguistic landscape of a commercial neighbourhood in Central Phnom Penh* *Journal of Multilingual and Multicultural Development* Vol. 33, No. 6, October 2012, pp. 553-567

Kiskowski, Maria A, Hancock, John F, Kenworthy, Anne K. (2009) *On the Use of Ripley's K-Function and Its Derivatives to Analyze Domain Size* *Biophys J.* (2009) Aug 19; 97(4): 1095–1103.

Landry, Rodrigue; Bourhis, Richard Y (1997) *Linguistic Landscape and Ethnolinguistic Vitality: An Empirical Study*. *Journal of Language and Social Psychology* March 1997 Vol. 16 No. 1 pp 23-49

Laerd Research Methodology--*Total Population Sampling*. Accessed on February 7, 2017. <http://dissertation.laerd.com/total-population-sampling.php>

- Lanza, Elizabeth; Woldemariam, Hirut (2014) *Language contact, agency and power in the linguistic landscape of two regional capitals of Ethiopia* International Journal of the Sociology of Language Vol. 2014, Issue 228, pp. 79-103 De Gruyter Mouton
- Lee, Jay & Kretzschmar, William A. Jr (1993) *Spatial analysis of linguistic data with GIS functions*, International Journal of Geographical Information Systems, 7:6, 541-560, DOI: 10.1080/02693799308901981
- Liao, Han-teng; Petzold, Thomas (2010) *Analysing Geo-linguistic Dynamics of the World Wide Web: The Use of Cartograms and Network Analysis to Understand Linguistic Development in Wikipedia* Journal of Cultural Science Vol 3, No 2 pp. 1-18
- Luebbering, Candicer. ; Kolivras, Korinen. ; Prisley, Stephenp.(2013) *Visualizing Linguistic Diversity Through Cartography and GIS* The Professional Geographer, 2013, Vol.65(4), p.580-593
- Milyaev, Sergey; Barinova, Olga; Novikova, Tatiana; Kohli, Pushmeet; Lempitsky, Victor (2013). "Image binarisation for end-to-end text understanding in natural images." (PDF). Document Analysis and Recognition (ICDAR) 2013. 12th International Conference
- Okabe, Atsuyuki and Yarnada, Ikuho (2001) *The K-Function Method on a Network and Its Computational Implementation* Geographical Analysis, Vol. 33, No. 3 July 2001
- Okabe, Atsuyuki, Satoh, Toshiaki, and Sugihara, Kokichi (2009) *A kernel density estimation method for networks, its computational method and a GIS-based tool* International Journal of Geographical Information Science Vol. 23, No. 1, January 2009, 7–32
- Okabe Atsuyuki, Okunuki K., Funamoto S. and Ishitomi T. 2002a. *A Toolbox for Spatial Analysis on a Network and its Software*. Proceedings of the 2nd International Conference on Geographical Information Science, Boulder, Colorado, USA.
- Okabe A., Sugihara K.(2012) *Spatial Analysis along Networks- Statistical and Computational Methods*. Volume 1 Wiley Chichester, pp. 135-136.
- Okabe, Atsuyuki; Okunuki, Kei-ichi; Shiode, Shino(2009) *SANET: A Toolbox for Spatial Analysis on a Network Version 3.4 – 121008* Center for Spatial Information Science University of Tokyo
- Okabe, A., Yomono, H. and Kitamura, M. (1995) *Statistical analysis of the distribution of points on a network*, Geographical Analysis, 27(2):152-175
- "Optical Character Recognition (OCR) – How it works". Nicomsoft.com. Retrieved 2013-06-16.
- O'Sullivan, D., & Unwin, D. J. (2002). *Geographic information analysis*. Hoboken, New Jersey: John Wiley.
- O' Sullivan, D., & Wong, D. W. S. (2007). *A surface-based approach to measuring spatial segregation*. *Geographic Analysis*, 39(2), 147–168.n Hall.
- Pati, P.B.; Ramakrishnan, A.G. (1987-05-29). Word Level Multi-script Identification. *Pattern Recognition Letters*, Vol. 29, pp. 1218 – 1229.
- Rafael, Eliezer; Shomany, Elana; Amara, Muhammad Hasan; Trumper-Hecht, Nira (2008) *Linguistic Landscape as Symbolic Construction of the Public Space: The Case of Israel* International Journal of Multilingualism Volume 3, Issue 1 pp. 7-30.

- Resig, John (2009-01-23). "John Resig – OCR and Neural Nets in JavaScript". Ejohn.org. Retrieved 2013-06-16.
- Ripley, B.D. (1976). *The Second-Order Analysis of Stationary Point Processes*. Journal of Applied Probability. **13**: 255–266.
- Ripley, D. (1976). *The Second-Order Analysis of Stationary Point Processes*. Journal of Applied Probability, pp. 13, 255-66.
- (1977). *Modelling Spatial Patterns*. Journal of the Royal Statistical Society, Series B, 39, 965-81.
- (1981). *Spatial Statistics*. Chichester: John Wiley.
- (1988). *Statistical Inference for Spatial Point Processes*. Cambridge: Cambridge University Press.
- Ripley B.D. *Modeling spatial patterns*. J. R. Stat. Soc. Series B Stat. Methodol. 1977;39:172–192
- SANET. A Spatial Analysis along Networks (Ver.4.1). Atsu Okabe, Kei-ichi Okunuki and SANET Team, Tokyo, Japan
- Schabenberger, O., & Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Boca Raton, Florida: Chapman & Hall/CRC.
- Sezgin, Mehmet; Sankur, Bulent (2004). "Survey over image thresholding techniques and quantitative performance evaluation" (PDF). Journal of Electronic imaging. 13 (1): 146.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman Hall.
- Smith, Ray (2007). "An Overview of the Tesseract OCR Engine"(PDF). Retrieved 2013-05-23.
- Smith, Tony E. *Notebook for Spatial Data Analysis*
- Spooner, Peter G.; Lunt, Ian D.; Okabe, Atsuyuki and Shiode, Shino. (2004) *Spatial analysis of roadside Acacia populations on a road network using the network K-function*. Landscape Ecology 19: 491–499.
- Tan, P. K. W. (2009). *Building names in Singapore: Multilingualism of a different kind*. In W. Ahrens, S. Embleton, & A. Lapierre (Eds.), *Names in multi-lingual, multi-cultural and multi-ethnic contact: Proceedings of the 23rd International Congress of Onomastic Sciences* (pp. 929–942). Toronto, Canada: York University.
- Torkington, K. (2009) *Exploring the linguistic landscape: The case of the 'Golden Triangle' in the Algarve, Portugal*. In Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching Vol. 3: Papers from LAEL PG 2008, ed. S. Disney, B. Forchtner, W. Ibrahim and N. Miller, 122_45. Lancaster: Lancaster University.
- Trier, Oeivind Due; Jain, Anil K. (1995). "Goal-directed evaluation of binarisation methods." (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 17 (12): 1191–1201
- UNDI.info: Malaysian Election Data. Last accessed on January 7, 2017. <http://undi.info/#/pahang>
- UNESCO World Report – *Investing in Cultural Diversity and Intercultural Dialogue* Accessed on January 1, 2017 <http://unesdoc.unesco.org/images/0018/001852/185202E.pdf>
- Van der Merwe, I.J. (1993) *The Urban Geolinguistics of Cape Town* GeoJournal, Vol. 31, No. 4 pp. 409-417
- Veselinova, L.N. (2009). *Studying the Multilingual City: A GIS-based Approach*. Journal of Multilingual and

Multicultural Development, Volume 30, pp. 145-65.

Wand, M.P. and Jones, M.C., 1995, *Kernel Smoothing* (London: Chapman & Hall/CRC).

Williams, Colin H., Merwe, Izak Van Der (1996) *Mapping the Multilingual City: A Research Agenda for Urban Geolinguistics* Journal of Multilingual and Multicultural Development, 1996, Vol.17(1), p.49-66 Taylor & Francis Group

Yamada, I., and J.-C. Thill. (2004). *Comparison of Planar and Network K-functions in Traffic Accident Analysis*. Journal of Transport Geography Vol. 12, 149-58.