Data Innovation

Metropolitan Sustainable Transportation

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

german cooperation DEUTSCHE ZUSAMMENARBEIT

Grab

PULSE LAB JAKARTA

GIZ Data Lab brings together thinkers and practitioners to promote the effective, fair and responsible use of digital data for sustainable development. The Lab applies an agile and experimental way of working to explore new data related trends and questions, and develop forward-looking solutions in GIZ partner countries. External partnerships - both when it comes to innovating and implementing - are key to the work of the Lab.

Pulse Lab Jakarta is a joint data innovation facility of the United Nations (Global Pulse) and the Government of Indonesia (via the Ministry of National Development Planning, Bappenas). The Lab employs a mixed-method approach, through which it harnesses alternative data sources and advanced data analytics methods to obtain actionable insights and applies human-centered design to ground-truth insights from its data analysis and research, providing evidence to inform policy makers.

Grab is the leading super app in Southeast Asia, providing everyday services that matter most to consumers. Today, the Grab app has been downloaded onto over 185 million mobile devices, giving users access to 9 million drivers, merchants and agents. Grab offers the widest range of on-demand transport services in the region, in addition to food, package, grocery delivery services, mobile payments and financial services across 349 cities in eight countries.

This report was produced by Pulse Lab Jakarta, co-authored in particular by Muhammad Rizal Khaefi (Junior Data Scientist) and Mellyana Frederika (Partnership and Advocacy Lead), with the guidance of Sriganesh Lokanathan (Data Innovation & Policy Lead) and support of Narawan Mahasarakam.

Table of Contents

Introduction

1 Context

Urbanization changes how the city operates. By 2050, it is expected that nearly 70% percent of the world's population will live in urban areas. Nearly 90% the increment of the urban population by 2050 is expected to be concentrated in Asia and Africa. In Southeast Asia, this phenomenon created mega urbans, such as Jabodetabek, Bangkok Metropolitan Region and Metro Manila. These cities are facing many transboundary challenges from water supply, wastewater to transport and environmental protection. Rapid urbanisation in this region has led to a large share of urban growth involving unplanned, unstructured expansion, with high rates of car use and a substantial proportion of people rely on personal vehicles to commute. Without adequate transport infrastructure, cities suffer the growing congestion on the road and there is a significant economic cost to this.

Transportation planning has been using data-driven analysis and modelling analysis to support policy decisions in this issue [1]. The data has traditionally informed transport planning has primarily been provided by manual surveys of people and their travel behaviour, for estimating travel demand, by manual mapping, service level and landscape surveys, for estimating transport supply, and by a mixture of manual and automated surveys of movements and transactions, for calibrating flows. However, data to develop and validate transport models has been limited by availability, frequency, or acquisition costs and time. Planners and policy makers are looking for ways to improve this.

2 Bangkok Metropolitan Transportation

In recent years, Southeast Asia cities such as Bangkok are rapidly expanding. Between the years 2000 and 2010, the capital of Thailand has grown from 1,900 square kilometers to 2,100 making it the fifth largest urban area in East Asia in 2010. Today, it is the second largest city in the region. However, this growth has not been aligned properly with Bangkoks land use and transport planning strategy, resulting in uncontrolled traffic growth and over reliance on private motorized vehicles. This unsustainable transport approach has increased traffic congestion, air pollution, massive investment in road infrastructure, high energy demand and GHG emissions, affecting the lives of millions of Bangkok's residents.

The Office of Transport and Traffic Policy and Planning (OTP) develops Extended Bangkok Urban Model (eBUM) as a transportation base model and using it as a tool to analyze and forecast transport and traffic situations. Initiated in 2016, the survey is done every five years with improvement and new data sources being fed into the model. New alternative data sources have potential to refine this model, validate assumptions and provide more responsive, near real time model.

3 The Project

The GIZ Data Lab is a platform that brings together thinkers and practitioners to promote the effective, fair and responsible use of digital data for sustainable development. The Lab applies an agile and experimental way working to explore new trends and develop forward-looking solutions in GIZ partner countries. Together with a network of partners, the Lab promotes the use of digital data to advance development and to support the delivery of innovative services by GIZ.

The GIZ Data Lab is a platform that brings together thinkers and practitioners to promote the effective, fair and responsible use of digital data for sustainable development. The Lab applies an agile and experimental way working to explore

new trends and develop forward-looking solutions in GIZ partner countries. Together with a network of partners, the Lab promotes the use of digital data to advance development and to support the delivery of innovative services by GIZ.

The Data Lab team up with Pulse Lab Jakarta, a joint initiative of the United Nations and the Government of Indonesia, to conduct a number experiment contributing to Sustainable Bangkok Metropolitan Transport. Pulse Lab Jakarta, established to explore the potential use of big data for public policy and social good, aim to harness big data and artificial intelligence responsibly as public good. Its mission is to accelerate the discovery and adoption of data innovation for sustainable development and humanitarian action.

This project is an experiment aimed at answering the question of how we can use non-traditional data sources to create more sustainable and inclusive transportation systems. The idea is to combine different data sources to improve the understanding of travel patterns for sustainable urban mobility planning.

4 Objective and Scope of Work

In recent years, the emergence of new technologies has greatly impacted personal mobility in various dimensions by modifying and/ or overcoming traditional travel barriers and constraints. The extensive use of smartphones by individuals has led innovators to develop app-based transportation services that efficiently link passengers to drivers within minutes. One of the most significant applications of ICT in transportation is ride-hailing a tech-driven adaptation of traditional street-hailing.It includes transportation network companies (TNC) offer methods of shared mobility that enable passengers to quickly book a ride directly with a vehicles owner using smartphone applications [2].

Ride-hailing transportation service allows passengers to request a ride in a real-time via smartphone application that links passengers to nearby drivers. This new form of on-demand transportation capitalizes on innovations like GPS chips to develop app-based, on-demand transportation that quickly and reliably connects riders and drivers. Both the passenger and the driver can use Global Position System (GPS) navigation during the trip, and the application guides the driver to shortcuts and less-congested roadways.

Ride-hailing applications are increasingly popular in Southeast Asian cities. Statista, a German online portal for statistics, stated that user penetration growth in the region is 16,8% in 2019 and is expected to hit 25,1% by 2023. In Bangkok alone, user penetration is 6,8% in 2019 and is expected to hit 11,7% by 2023 [3]. This growth leads to significant increase in the data being generated from ride-hailing applications related to mobility patterns in the region. GrabTaxi Holdings Pte. Ltd. (branded as simply Grab) is a Southeast Asia-based technology company that offers ride-hailing, ride-sharing and logistics services through its app, covering Singapore, Cambodia, Indonesia, Malaysia, Philippines, Vietnam, Thailand, and Myanmar. Grab joined this collaboration through their partnership with Pulse Lab Jakarta in promoting mobility and transport accessibility in a responsible and ethical manner. Grab as data holder provides the main alternative data sets in this experimentation and the scope of the experimentation is as follows:

Based on this opportunity, the scope of the experimentation is as follows:

1) Exploring the use of ride-hailing data to inform macroscopic traffic flow modeling.
2) Exploring the use of ride-hailing data to develop road speed profile.
3) Exploring the use of ride-hailing data to nowcast location, time and patterns of traffic to inform the formulation of a congestion charge.
4) Proof of concept to use ride-sharing data as a proxy for measuring population activity patterns to calculate population-weighted exposure to air pollution on a city-wide scale.

Note: This is a different scope of work with a plan. The initial plan to conduct experiments to infer road quality using ride-hailing data is changed to develop road speed profiles as requested by GIZ.

5 DATASETS

5.1 Ride-hailing Data

We use anonymous ride-hailing data sets collected from September 2018 to December 2018 and from March 2019 to April 2019 to capture the different patterns between common days and uncommon days. To represent uncommon days, we select days during Songkran Festival, the Thai New Years national holiday on the 13th April every year and can be extended to 15th April or more. The size of this data set is over 4 billion measurements.

Ride-hailing data used in this experiment is trajectory data. It is collected through the driver's Grab Application and uploaded to the data management center by Grab Application. The trajectory data is reported at a 30-s interval, including real-time information as described in Table I such as longitude and latitude (including altitude), instantaneous speed, moving direction (360 degrees) and timestamp. Other ride-hailing data generated by accelerometer and gyroscope at high spatio-temporal resolutions.

5.2 Transportation Data

The Office of Transport and Traffic Policy Planning of the Government of Thailand (OTP) uses a trip-based transportation model called Extended Bangkok Transport Model. The model has been used to analyse and forecast transport and traffic situations in Bangkok in the past 15 years.

From this model, OTP provided access to 2017 congestion data and its estimation for 2027 and a road map network including public transportation network data. In addition, we received access to congestion data and 2027 estimation from OTP. The congestion data is an integrated part of the extended Bangkok Urban Model. Complete lists of the transportation data used in this research are shown in Table II.

5.3 Environment Data

Pollution Control Department of Government of Thailand (PCD) provided historical air pollution data collected from 12 official sensors in Bangkok. In addition, we used open data sources to collect other environment data as listed in Table III.

5.4 Demography Data

We have limited access to socio-demography and land use data of Bangkok Metropolitan Area. However, we found basic information on population, city administration, and important landmarks as proxy to detail datasets needed as local characteristics to our model. Full lists demographic data used in this study are shown in Table IV.

TABLE I: Ride hailing metadata

No	Column Name	Description	Example of value
1	driver id	Anonymized driver id (64 characters)	ec79db0575ad620621ede0d3 d36740c918731a5f0b4990d4 961a9552f1cef6a4
2	timestamp	Date and time (yyyy-mm-dd)	2019-01-01
3	latitude	Geographic coordinate of east-west position on the Earth's surface (degree)	-6.25
4	longitude	Geographic coordinate of north-south position on the Earth's surface (degree)	108.25
5	altitude	Height relative with sea level (m)	10
6	bearing	Direction relative to the true north (degree)	100
7	speed	Moving speed (m/s)	20
8	gyroscope data	Device orientation and angular velocity x,y,z	1,-1,2
9	accelerometer data	Device acceleration forces x,y,z	2,-1,3

TABLE II: Transportation metadata

No	Data	Details	Source	Year
1	Roadway and transportation network	• Road length • Road type • Number of lanes	Bangkok OTP	2011
2	Public transportation network	• Land based (BTS (Skytrain), MRT, Bus) • River based (Chao Phraya Expressway, Khlong Saen Saep boat express)	Bangkok OTP	2019
3	Historical congestion data	Estimated congestion level developed by eBUM model	Bangkok OTP	2017
4	Road construction event	• Location and time of construction work • Lane direction	Government official release and news	2018-2019

TABLE III: Environment metadata

No	Data	Details	Source	Year
1	Historical air pollution	Air quality level on a daily basis.	PCD	2018-2019
2	Aerosol optical depth (MCD19A2)	Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm-based Level-2 gridded (L2G) AOD.	US National Aeronautics and Space Agency (NASA)	2018-2019
3	Enhanced vegetation indices (MOD13A3)	Global MOD13A3 data are provided every month at 1-kilometer spatial resolution as a gridded Level-3 product in the Sinusoidal projection	NASA	2019
4	Digital elevation model (SRTM)	Shuttle Radar Topography Mission (SRTM) provide complete digital elevation data for global coverage	United States Geographical Survey Agency	2014
5	Air temperature	Daily aggregate of air temperatures from ground sensors measurements	US National Oceanic and Atmospheric Administration (NOAA)	2018-2019

TABLE IV: Demography metadata

No	Data	Details	Source	Year
1	Population	Number of population at sub-district level	Bangkok Office of Statistics	2015
2	Population	Population estimation at 100m x 100m grid	Worldpop	2014
3	City Administration Map	Administration boundary down to sub-district level	OTP	2011
4	Landmark	• Type of landmark (commercial, residential, office, etc) • Geo-location	OTP	2011

Data for Mobility

The Office of Transport and Traffic Policy and Planning (OTP) developed and used a trip based transportation model called Extended Bangkok Urban Model (eBUM). This is the most complete and up-to-date transportation model developed in Bangkok. It is a tool to analyse and forecast transport situations as a result of changes in transport network in the study area. The model is also used to test traffic management measures proposed by responsible agencies. An eBUM consists of fundamental data which is updated through an expensive survey and this is one important issue in using eBUM. Another challenge is in the model calibration and validation process.

We see opportunity to use this experiment to explore ways that potentially could tackle some challenges eBum is currently facing by asking the following question: a) can ride-hailing data be used to validate pattern of traffic flow; b) is it possible to use ride-hailing data to develop road speed profile, to compliment one develop through eBUM; c) is is plausible to harness ride-hailing data to predict traffic congestion. A more ambitious question but nevertheless important is can ride-hailing data be used to quantify the population exposure to air pollution. There were precedents to track the pollution exposure using mobile-device-based mobility patterns to consider spatial and temporal aspects of the exposure estimates.

The next section explains the the research in this following order: a) Macroscopic Traffic Flow Modelling; b) Road Speed Profiling; c) Traffic Congestion Nowcasting; and d) Quantifying Popultion Exposure to Air Pollution. This section describe the method employ by each research and key results.

As initial attempt, the research shows potential of using ride-hailing data to compliment exising transportation model used by OTP. However, this result has to be taken carefully because the different degree of potential and the extend in which the experiment able to answer the research questions. Therefore, in conclusion section, we wrote our reflections and recommendation for further works.

Macroscopic Traffic Flow Modelling

1 Summary

Macroscopic traffic flows are one of the basic insights needed for transportation planning and modelling. It allows planners and modelers to important properties related to traffic flow including how they are generated. Macroscopic traffic flows are useful for short-term forecasting providing the ability to understand and calculate, amongst others, average travel times, mean fuel consumption, etc. Traditional methods for developing macroscopic traffic flows depend on data collected through field surveys and observations, which are time consuming to conduct, but also have limited spatio-temporal coverage, and can be fairly subjective. Commercial products for getting those information, including Google Maps and Waze are popular and widely used but considered costly for city wide analysis. A trip distribution model is an alternative approach using transportation theory as a proxy to infer the traffic flows. The model promises lower cost, the need for fewer data, and huge spatio-temporal coverage but suffers from localization constraints, which need city-specific calibration. This research explore the possibility of using Grab data to assess the performance of four trip distribution models to proxy traffic flows in Bangkok. These four models include: i) a gravity model with exponential decay, ii) a gravity model with power law distance decay, iii) a radiation model, and iv) a radiation-extended model. Results are calibrated with actual traffic activity inferred from Grab data to prove the feasibility. The radiation model shows better correlation score ($\rho = 0.5$) with inferred actual traffic flows as compared to the others.

2 Data and Features

1) Population count by districts pop_{otp}, from Bangkok's Office of Transport Planning and Worldpop
2) Administrative boundary, from Bangkok's Office of Transport Planning
3) August,December 2018 and March-April 2019 Trajectory data, from Grab

3 Methods

The purpose of the trip distribution models is to split the total number of trips N in order to generate a trip table $\tilde{T}$ of the estimated number of trips from each census area to every other. For simplicity, for this preliminary work we restrict our analyses to trips from home to work, without also considering the return journey. N is equivalent to the number of unique commuters. The trip distribution depends both on the characteristics of the census units and the way they are spatially distributed, as well as on the level of constraints required by the model. Therefore, to fairly compare different trip distribution modeling approaches we have to consider separately the law used to calculate the probability to observe a trip between two census units, called trip distribution law, and the trip distribution model used to generate the trip allocations from this law. The four trip distribution models tested [4] are the following:

3.1 Gravity laws

In the simplest form of the gravity model based approach, the probability of commuting between two units i and j is proportional to the product of the origin population m_i and destination population m_j, and inversely proportional to the travel cost between the two units:

$$p_{ij} = m_i m_j f(d_{ij}), i \neq j$$

The travel cost between i and j is modeled with an exponential distance decay function,

$$f(d_{ij}) = e^{-\beta d_{ij}}$$

and a power distance decay function,

$$f(d_{ij}) = d_{ij}^{-\beta}$$

Barthelemy et.al [5] found that the form of the distance decay function can change according based on the dataset. Therefore, both the exponential and the power forms are considered in this study. In both cases, the importance of the distance in commuting choices is adjusted with a parameter β with observed data.

3.2 Radiation laws

In the intervening opportunity approach, the probability of commuting between two units i and j is proportional to the origin population m_i and to the conditional probability that a commuter living in unit i with population m_i is attracted to unit j with population m_j, given that there are s_{ij} job opportunities in between. The conditional probability $\mathbb{P}(1|m_i, m_j, s_{ij})$ needs to be normalized to ensure that all the trips end in the region of interest.

$$p_{ij} = m_i \frac{\mathbb{P}(1|m_i, m_j, s_{ij})}{\sum_{k=1}^{n} \mathbb{P}(1|m_i, m_k, s_{ik})}, i \neq j$$

Simini et al. [6] proposed diffusion model where each individual living in an unit i has a certain probability of being absorbed by another unit j according to the spatial distribution of opportunities, called radiation model. The original radiation model is free of parameters and, therefore, it does not require calibration. The conditional probability is expressed as:

$$\mathbb{P}(1|m_i, m_j, s_{ij}) = \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}$$

An extended radiation model has been proposed by Yang et al. [4]. In this extended version, the probability $\mathbb{P}(1|m_i, m_j, s_{ij})$ is derived under the survival analysis framework introducing a parameter α to control the effect of the number of job opportunities between the source and the destination on the job selection,

$$\mathbb{P}(1|m_i, m_j, s_{ij}) = \frac{\left[(m_i + m_j + s_{ij})^{\alpha} - (m_i + s_{ij})^{\alpha}\right](m_i^{\alpha} + 1)}{\left[(m_i + s_{ij})^{\alpha} + 1\right]\left[(m_i + m_j + s_{ij})^{\alpha} + 1\right]}$$

After the description of the probabilistic laws, the next step is to materialize the people commuting. The purpose is to generate the commuting network $\tilde{T}$ by drawing at random N trips from the trip distribution law p_{ij} In the absence of empirical commuting data, we consider only the unconstrained approach to randomly sample the N trips from the multinomial distribution,

$$\mathcal{M}\left(N, (p_{ij})_{1 \leq i,j \leq n}\right)$$

4 Results

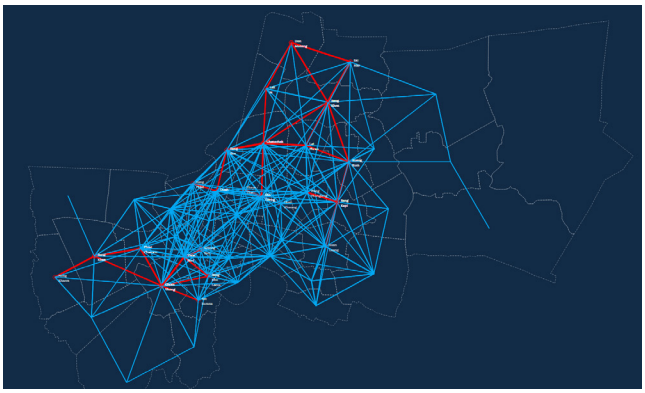
Table I and Figure 1 show the four trip distribution models along with the results extracted from Grab data. Both Gravity Models show dominance of the bottom left and top right regions. The four highest traffic flow districts based on Exponential Spatial are Sathon, Thon Buri, and Wang Thonglang. The Power Law variant's four highest traffic flow districts are Sampanthawong, Sai Mai, and Pom Prap districts. Pearson correlation is calculated for each to measure the similarity with inferred actual traffic flows. Both gravity models achieve relatively low scores with gravity exponential model having $\rho = 0.4$ and gravity power law having $\rho = 0.1$, which indicates dissimilarity. The low scores are explained as only a handful of top-10 districts on both gravity model are found in the top-10 of inferred actual traffic flows. Both gravity models also failed to capture high traffic flows in Bangkok's central regions.

Different to the Gravity Model, the two Radiation Models show relatively even traffic flow distributions in all regions. The dominant regions are the following: Ratchathewi, Thon Buri, and Wang Thonglang for Standard Radiation Model and Wang Thonglang, Thon Buri, Khlong Toei for the Radiation-Extended Model. Standard Radiation Model gives $\rho = 0.5$ correlation scores(which is highest score compared with other models), and the extended variant only achieved $\rho = 0.2$ score. Standard Radiation Model's high scores suggest it is best suited to identify high traffic flow in central Bangkok and important traffic hubs.

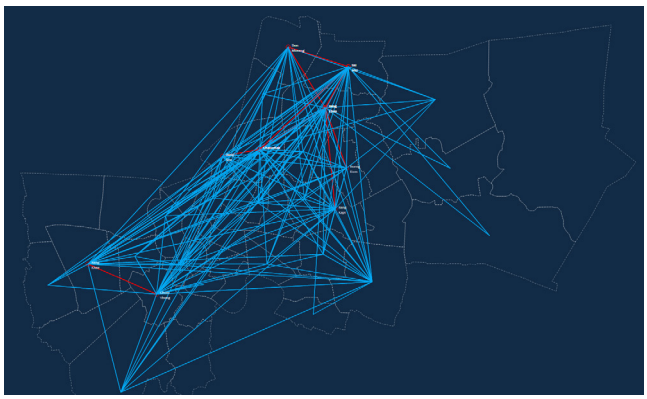
This promising result should be treated cautiously. All models show large deviation with actual traffic flow ($adj.r^2 = 0.25$) inferred from Grab data which indicates noticeable numerical gap.

TABLE I: Top-10 Highest Traffic Flows by Districts

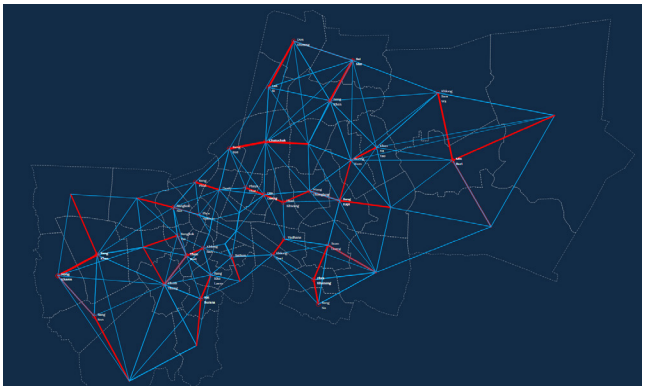
No	Inferred from Grab		Gravity - Exponential		Gravity - Power		Radiation		Radiation - Extended	
	District	% flows	District	% flows	District	% flows	District	% flows	District	% flows
1	Chatuchak	6.2	Sathon	3.2	Sampanthawong	3.3	Ratchathewi	3.7	Wang Thonglang	3.1
2	Ratchathewi	6.1	Thon Buri	3.2	Sai Mai	2.7	Thon Buri	3.7	Thon Buri	2.9
3	Pathum Wan	5.8	Wang Thonglang	3	Pom Prap S.P.	2.6	Wang Thonglang	3.7	Khlong Toei	2.8
4	Khlong Toei	5.6	Suan Luang	2.9	Lat Phrao	2.6	Sathon	3.6	Sathon	2.8
5	Huai Khwang	4.9	Khlong Toei	2.9	Don Mueang	2.5	Phaya Thai	3.2	Suan Luang	2.5
6	Vadhana	4.8	Taling Chan	2.8	Taling Chan	2.4	Din Daeng	3.1	Rat Burana	2.5
7	Bang Rak	3.9	Yan Nawa	2.7	Phra Nakhon	2.4	Pom Prap S.P.	3	Sai Mai	2.5
8	Din Daeng	3.6	Ratchathewi	2.6	Chom Thong	2.4	Vadhana	3	Saphan Sung	2.5
9	Bang Kapi	3.4	Phasi Charoen	2.6	Khan Na Yao	2.4	Khlong Toei	3	Yan Nawa	2.5
10	Suan Luang	3	Phra Nakhon	2.5	Chatuchak	2.3	Suan Luang	2.9	Ratchathewi	2.4



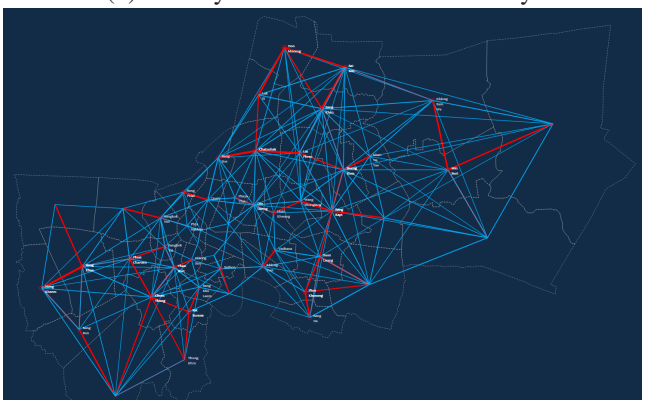
(a) Gravity Model - Exponential Decay



(b) Gravity Model - Power Law Decay



(c) Radiation Model



(d) Radiation Extended Model

Fig. 1: Comparison of generated traffic flows in Bangkok

Road Speed Profiling

1 Summary

Road-speed profiling refers to the process of computing the expected, median, upper-bound and lower-bound (within a confidence interval) speeds of road segments in a city at a given time of the day. [7] Road speed profiling is used to measure point-to-point travel time estimation, build advanced traffic information and management system that could eventually help ease congestion given the constraints in term of building new infrastructure. Aggregating road speed is challenging task because vehicle speeds are inherently stochastic. Traditionally, road-speed profile is calculated based on field survey and observation data, with limited spatio-temporal coverage. This research explores ways to improve road-speed profile accuracy by developing road-speed profile for 9 major roads in Bangkok in 15 minutes intervals in a day using one month ride-hailing data provided by Grab. The road profile shows promising result that reflect mobility pattern between normal days and major event day.

2 Data and Features

1) August, December 2018 and March-April 2019 Trajectory Data
   - Location
   - Timestamp
   - Speed
2) Digitized Road Network
   - Road link/segment
   - Road length
   - Road type

3 Methods

The data used represents real speed measurements from smartphone sensors from the period of September 2018 to December 2018 and March to April 2019. This data provide accurate and representative picture of the actual driving conditions across the Bangkok road network as shown in Figure 1.

A $Trajectory$ of a moving grab driver is a sequence of discrete points in geographical space, articulated as geo-locations with corresponding timestamps and speed profile, i.e.,

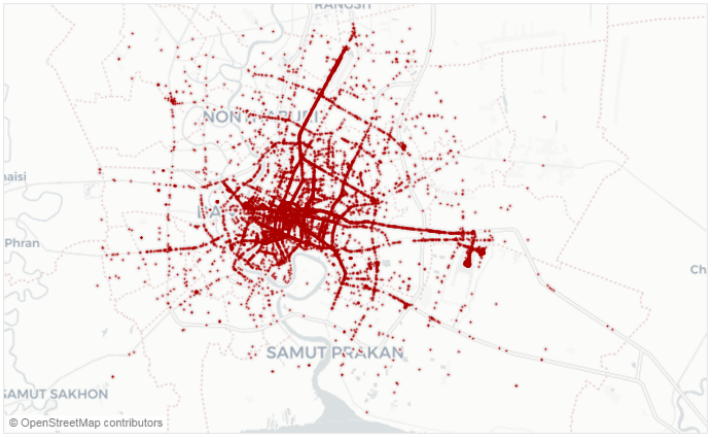
$$Trajectory = < p_1, s_1, t_1 >, < p_2, s_2, t_2 >, ..., < p_n, s_n, t_n >,$$

where each element $< p_i, s_i, t_i >$ indicates a driver is at location p_1 at timestamp t_1 with speed s_1 m/s. Furthermore, elements are sorted by timestamps, i.e., $t_j < t_k$ if $1 \leq j < k$.

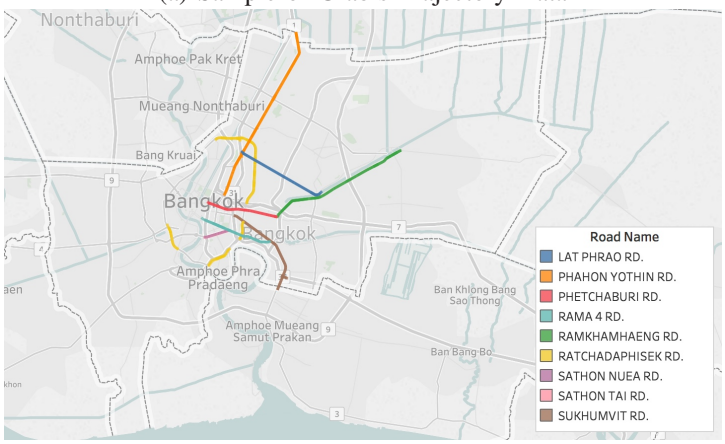
The spatial reference for this research is the digitized road network provided by Bangkok's Office of Transport Planning. The road network consists of road areas, a smaller segment called road link, and also additional information, such as length and road type. For every road link there will be lanes, which we classified as inbound and outbound. Since the trajectory data doesn't have specific information to indicate which lane is being utilized, we inferred lane directions are inferred using heading information and k-means clustering algorithms. Prior to the focus area for this study, only primary and important roads in Bangkok are chosen.

3.1 Preprocessing

GPS satellites broadcast their signals in space with a certain accuracy. This depends on additional factors, including satellite geometry, signal blockage, atmospheric conditions, and receiver design features affect the quality of data. GPS-enabled smartphones are typically accurate to within a 4.9 m (16 ft.) radius under open sky [8]. However, their accuracy worsens near buildings, bridges, and trees. The GPS data used for this research is mostly taken on open roads with a few exceptions such as underpass roads. Ten roads were selected: Phahon Yothin, Ratchadaphisek, Sathon, Sukhumvit, Rama 4, Phetchaburi, Ramkhamhaeng, Lat Phrao, and



(a) Sample of Grab's Trajectory Data



(b) Road sample

Fig. 1: Visualization of Data used

Chaeng Watthana. In order to reduces the uncertainty, a 50m buffer was chosen. Few speed observation outliers were found (e.g. some had speeds of about 1000 km/h), which are then removed using the Interquartile range (IQR) method.

3.2 Speed Profiling Calculation

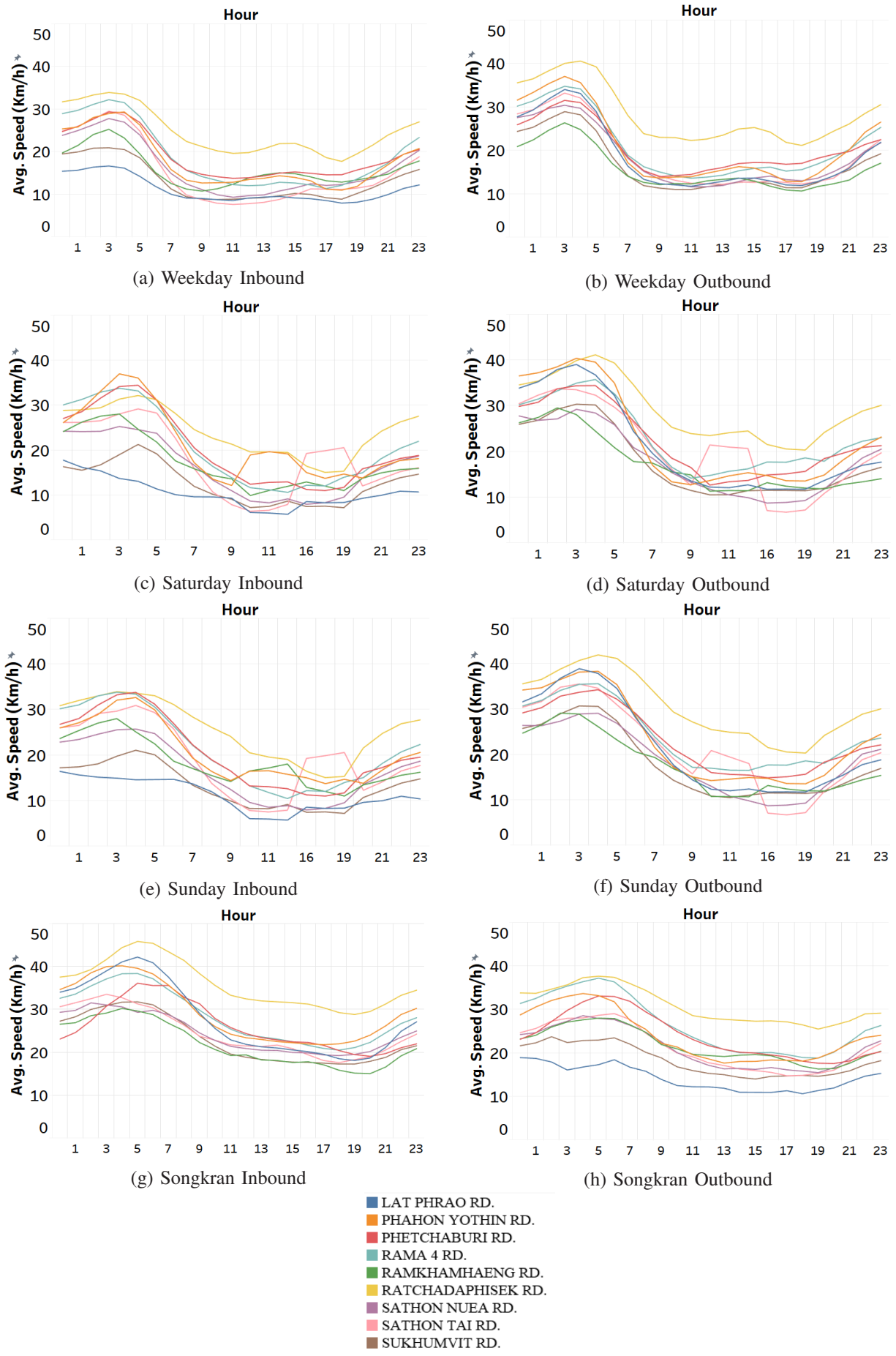
The average speed on a road link can be calculated as the moving average of probe speeds over a period p:

$$\overline{v}_{i,p} = \frac{1}{p}\sum_{j=1}^{i+p} v_j$$

where p is the duration f the time period p, i is the start of the period, and v_j is the average probe speed in time slot j. [9] Based on consultation we choose 15 minutes time period in this study.

4 Results

Figure 2 shows the speed graph during weekday, weekend and Songkran Festival at different lane direction relative to city center (inbound: going to, outbound: leaving). Temporal patterns during weekend and weekday are similar, these are normal days. Moving out from city center is faster with highest average speed at 3 am to 4. The pattern changes during Songkran Festival. The speed to move from the city center to outside is slower compare to normal days, indicates more people leave the city, with highest speed at 5 am to 7 am. The difference between weekday and weekend is in lowest speed time, from 7 am to 7 pm during weekday and 4 pm to 7 pm during weekend. There is no significant variation regarding spatial patterns, Rama 4 is where the speed is the lowest and the Ratchadaphisek is road with highest speed profile.



(a) Weekday Inbound

(b) Weekday Outbound

(c) Saturday Inbound

(d) Saturday Outbound

(e) Sunday Inbound

(f) Sunday Outbound

(g) Songkran Inbound

(h) Songkran Outbound

Fig. 2: Speed graph comparison during weekday, saturday, sunday, and songkran festival

Traffic Congestion Nowcasting

1 Summary

Traffic congestion is on of the most important problem domains for transportation planning and management. The causes can be complex and sometime random. The prediction of urban traffic congestion has emerged as one of the most important research topics of advanced transportation system. In the field of traffic flow forecasting, a number of models and methodologies have been put forward for the improvement of the existing model. This research aim to nowcast traffic congestion using the following predictors: historical congestion, road-speed profile, population count, district characteristics and time seasonality. The model able to capture hourly seasonality and daily seasonality with limitation during heavy congestion. In addition, this research explore the possibility of using the result to improve Extended Bangkok Urban Model (eBUM) by comparing the result of both models. It shows mixed results among different roads which indicate the need to conduct further research that use digital data for both models.

2 Data and Features

1) August, December 2018 and March-April 2019 Trajectory Data
   - Location
   - Timestamp
   - Speed
2) Digitized Road Network
   - Road link/segment
   - Road length
   - Road type
3) City Landmark (as proxy to land use)
   - Residential density
   - Commerce density
   - Office density
   - Industrial density
   - Tourism density
4) Bangkok Administrative Boundary
   - Districts area
5) Population
   - Population density by districts

3 Methods

3.1 Preprocessing and Feature extraction

The analysis is conducted using processed data from road-speed profile describe in previous section. Density of population and city landmark calculated by summing all units (e.x. population, point of interests, or office buildings) within district divide by the district's area size.

$$density = \frac{\sum units}{area_{district}}$$

3.2 Congestion level calculation

In order to calculate congestion level, free flow speed has to be determined. Free flow speed is considered as the maximum measured average speed over the period:

$$\overline{v}_{free} = \max_{j \in [t_0, t]} v_j$$

The congestion level is defined by ratio of the average speed during the most congested period p with respect to the maximum speed:

$$c_p = \frac{\overline{v}_p}{v_{free}}$$

3.3 Model development and validation

For a given data set with n observations, consisting of congestion level c_p with district and road characteristics m at time window period p, formulated by

$$\mathcal{D} = \{(x_{ijt}, c_p)\}(|\mathcal{D}| = n, x_{ijt} \in \Re^{M_i, M_j, E_t}, c_p \in \Re)$$
$$M =< pop, landuse, road_{profile}, districts_{area} >$$

This is follow by development of a tree ensemble model [10] in the form of K additive functions, which describe relations between congestion level with district characteristics, time, and the external factor,

$$c_p = \phi(x_i) = \sum_{k=1}^{k} f_k(x_i), f_k \in \mathcal{F}$$

Here q represents the structure of each tree that maps an example to the corresponding leaf index. T is the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and leaf weights w.

$$\mathcal{F} = \{f(x) = w_{q(x)}\}(q : \Re^{M_i, M_j, E_t} \rightarrow T, w \in \Re^T)$$

To find the best model, κ additive function and Ω model complexity is calculated to minimize the objective function,

$$\mathcal{L}(\phi) = \sum_{ijt} l(\widehat{c_p}, c_p) + \sum_{\kappa} \Omega(f_k)$$

Convex loss function l measures the difference between congestion level prediction $\widehat{c_p}$ and actual congestion level c_p.

In order to measure model performances,10-fold cross validation approach used for splitting the data set into training and testing.

3.4 Comparison with eBum model

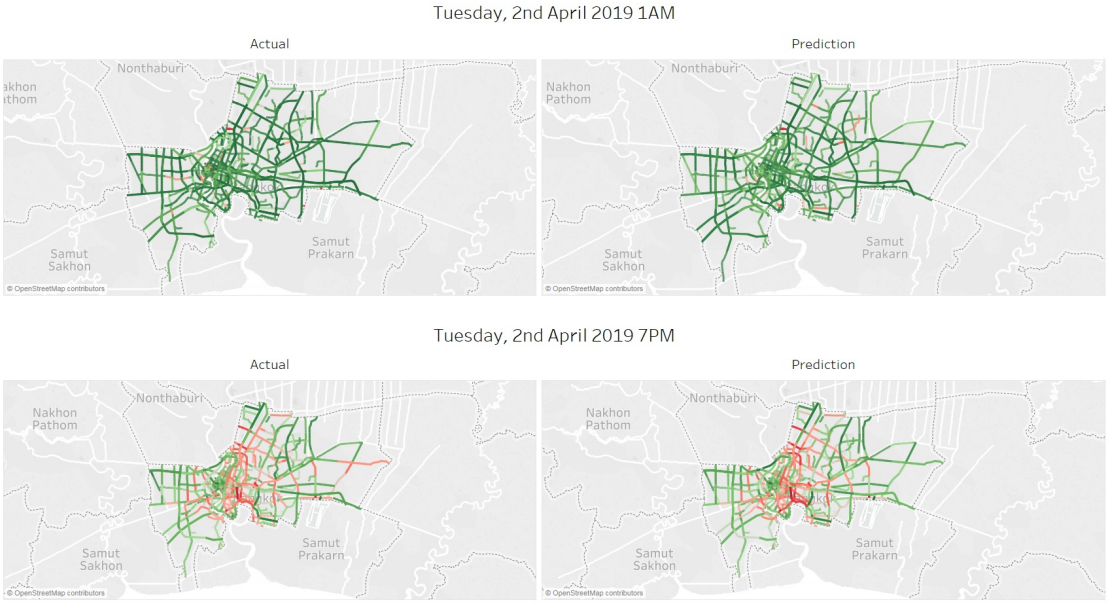
The last stage is to compare congestion level inferred from nowcast model with congestion level from eBum in 2017 as shown in Figure 3a. Unfortunately, congestion level from eBum model is not available in digital version and therefore this comparison is performed manually, with on-the-field human observation. The comparison is done by using snapshot of congestion level from both models, focusing on area shown in eBum model and mirroring the spatial resolution of eBum model snapshot. The comparison is conducted by selected Bangkok's residents who conducted on-the-field congestion observation indicated by different colour code of both model output, and assess the color similarity range from 1 (totally different) to 10 (identical).

This comparison is conducted in limited area. This area is selected based on spatial resolution of eBum's model. Six districts are selected, they are Phaya Thai, Dusit, Bang Plat, Ratchathewi, Pathumwan, and Bangkok Noi. Selected road are Phahon Yothin, Ratchadaphisek, Sathon, Sukhumvit, Rama 4, Phetchaburi, Ramkhamhaeng, Lat Phrao, and Ram Inthra.

4 Results

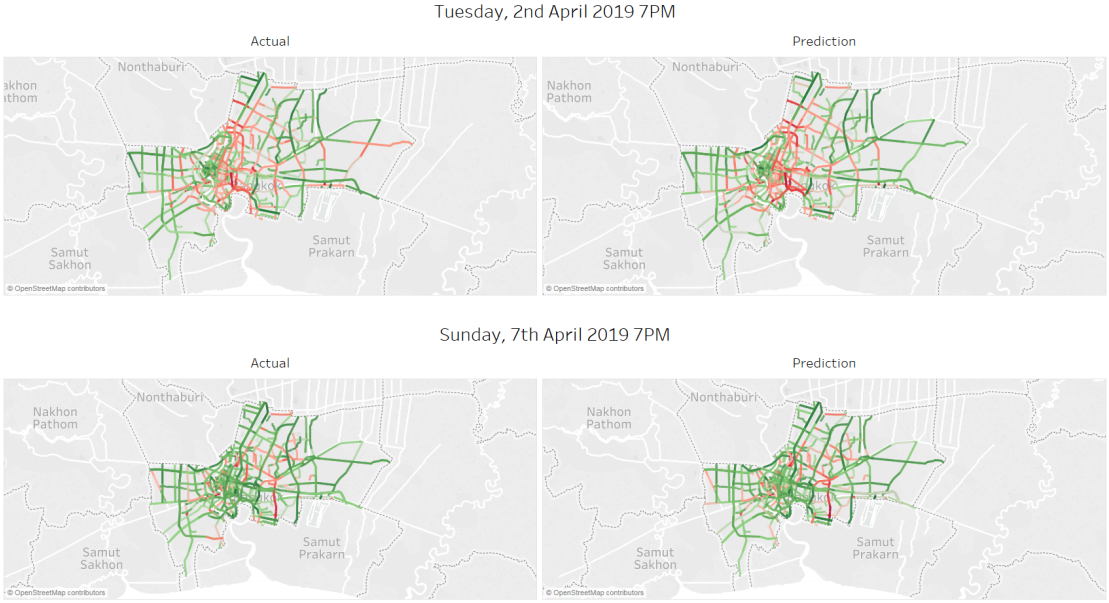
A set of predictor features ranging from historical congestion, road profile, population count, district characteristics and time seasonality is used to inform the development of nowcast congestion model. This research develops model to predict congestion level during week of 1st to 7th April 2019. Current best model could nowcast congestion with $rmse = 0.1$, $r = 0.7$, and $r^2 = 0.5$. In general, the model able to capture hourly seasonality (eg. able to predict congestion during peak hour) and daily seasonality (ex. able to predict less congestion during weekend) as shown in Figure 1. Figure 2 shows kernel density estimation plot of actual and predicted congestion. It shows similar proportion for congestion below 0.6 but not for the above levels which tends to underestimate heavy congestion ($0.6 \leq c_p \leq 0.9$) and overestimate the severe congestion ($c_p = 1$).

Table Ia and Ib show similarity scores gathered from respondents. At the district groups, most respondents indicated agreement with the idea that Phaya Thai, Bang Plat, and Pathumwan have good similarity in both model ($Median = \{5, 6\}.IQR = \{0, 1\}$). Opinion seems to be divided in the rest of the districts as equal number of respondents expressed strong disagreement or disagreement which indicates no consensus ($IQR \geq 2$). Looking at the road groups, consensus reached at the several roads ($IQR = \{0, 1\}$), some of them have relatively high similarity scores $Median = 6$ like Ramkamhaeng, Phetchaburi, and Rama 4. On the other hand, some roads are perceived different $Median = 4$ like Ratchadaphisek and Lat Phrao.



(a) Hourly seasonality

seasonality.png



(b) Daily seasonality

Fig. 1: Congestion nowcasting results

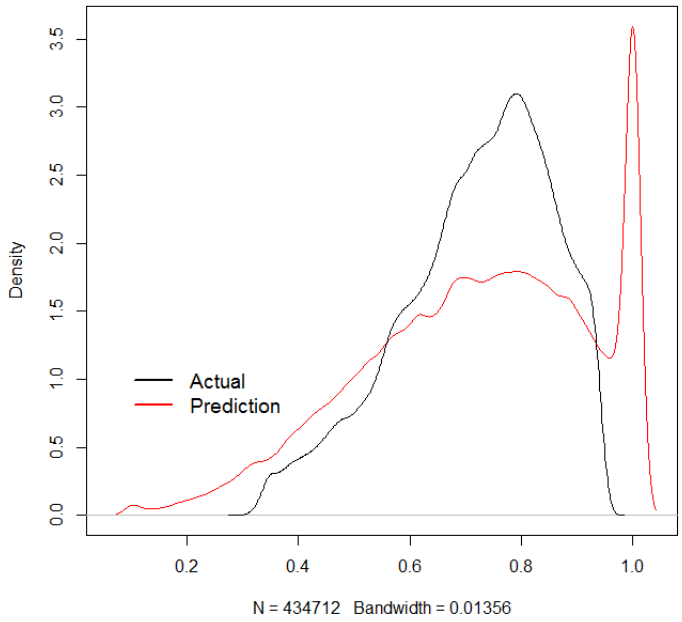


Fig. 2: Kernel density plot of Actual and Predicted Congestion

No	District Name	Similarity Score (1-10)					Median	IQR
		#1	#2	#3	#4	#5		
1	Phaya Thai	5	6	6	10	6	6	0
2	Dusit	7	4	5	10	6	5	2
3	Bang Plat	6	5	6	10	7	6	1
4	Ratchathewi	3	6	7	1	4	3	3
5	Pathumwan	6	5	6	1	5	5	1
6	Bangkok Noi	4	4	6	1	6	4	2

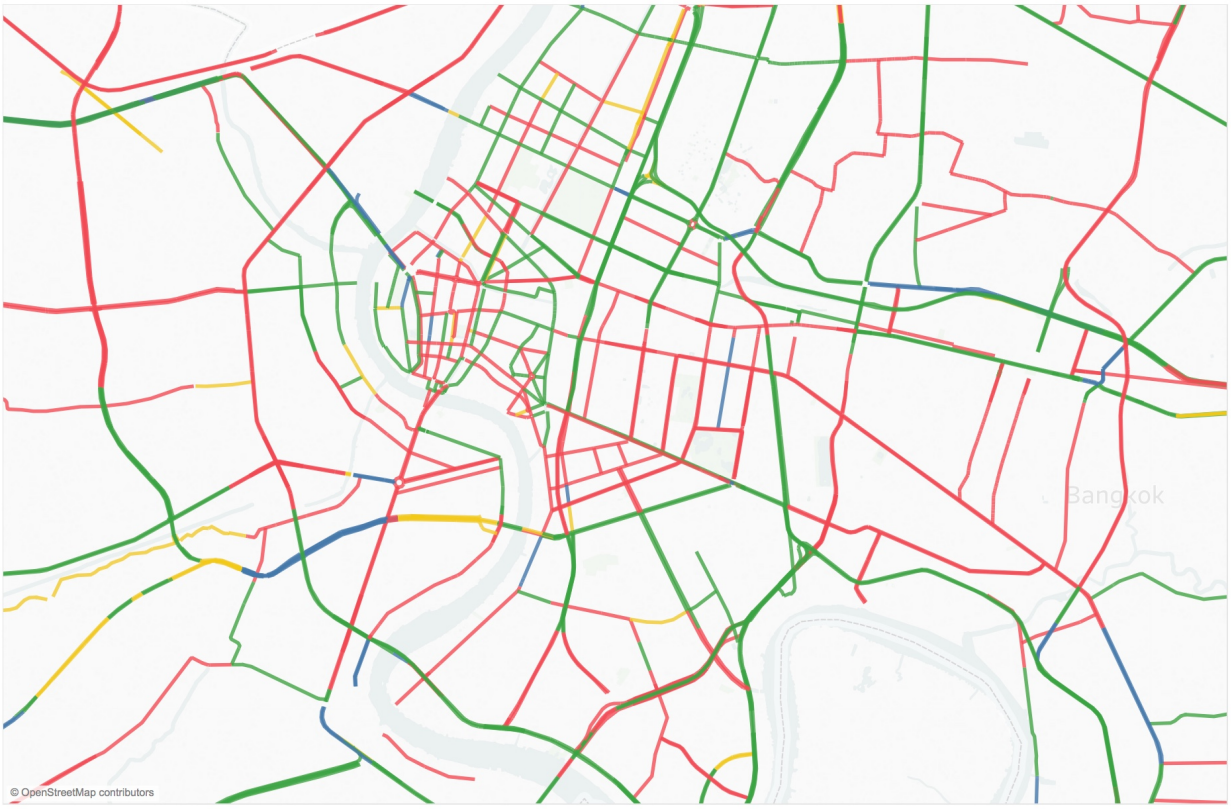
(a) By Districts

No	Road Name	Similarity Score (1-10)					Median	IQR
		#1	#2	#3	#4	#5		
1	Phahon Yothin	4	6	7	7	5	5	2
2	Ratchadaphisek	4	4	5	5	5	4	1
3	Sathon	4	3	6	5	6	4	2
4	Sukhumvit	7	4	5	5	7	5	2
5	Rama 4	8	6	7	6	7	6	1
6	Phetchaburi	7	7	6	6	6	6	1
7	Ramkamhaeng	2	6	6	6	6	6	0
8	Lat Phrao	4	3	5	5	6	4	1
9	Ram Inthra	2	6	6	6	5	5	1

(b) By Roads

TABLE I: Respondent scores

(a) Ebum model



(b) Congestion model based Grab data

Fig. 3: Comparison of Congestion model

Quantifying Population Exposure to Air Pollution

1 Summary

Accurate estimates of human exposure to inhaled air pollutants are necessary for a realistic appraisal of the risks these pollutants pose and for the design and implementation of strategies to control and limit those risks. This estimation, except in occupational setting, are usually based on measurements of pollutants concentrations in outside air, recorded with outdoor-fixed site monitors. From public health perspective, it is important to determine the population exposure - the aggregate exposure for a specified group of people. This research aim to explore ways of harnessing four months observation data to quantifying population exposure to air pollution. Using Land Use Regression (LUR), this research infer daily air quality in Bangkok at 1kmx1km level. Bangkok Metropolitan has 13 official ground air quality sensors, although limited in number and coverage, data from these sensors are sufficient for preliminary validation. The best model shows $r^2 = 0.6$ inference performance.

2 Data and Features

1) Air Quality Level AQI
2) Traffic Congestion (reference: Traffic Congestion Nowcasting)
3) Digitized Road Network
4) Aerosol Optical Depth (AOD) at 047 and 055 micron aod_{047}, aod_{055}
5) Open spaces
   - Enhanced Vegetation Indeces evi
   - Normalized Difference Vegetation Indeces $ndvi$
6) Digital Elevation Model dem
7) Air temperature $temp$
8) Population density pop

3 Methods

3.1 Preprocessing and Feature extraction

Based on Bangkok Metropolitan administrative boundary, a reference grid cells of 1km x 1km are created. For each predictors (congestion, main road, AOD, open spaces, etc), a spatial and temporal aggregation performed by calculate predictor observation j as many n inside grid i at the day d,

$$\overline{predictors}_{i,d} = \frac{1}{n}\sum_{j=1}^{n} predictors_j$$

$$predictors =< congestion, mainroad, aod_{047}, aod_{055}, evi, ndvi, dem, temp, pop >$$

As the first step, AQI levels data are investigated to assess quality of its measurements. Correlation analysis are conducted for each ground sensor sites. Ground sensor site with low correlation score ($r \leq 0.2$) were removed from next iteration. Land use regression with 3 steps explained below performed [11; 12].

3.2 Developing model to infers AQI from predictors

Model to infers AQI level relationship with available predictors (referred as type 1) and additional data are developed by creating training data set with n observations, consisting of AQI level AQI_{ij} and type 1 predictors at a grid cell i on a day d, formulated by

$$\mathcal{D} = \{(x_{id}, aqi_{id})\}(|\mathcal{D}| = n, x_{id} \in \Re^{predictors^{type1}}, aqi_{id} \in \Re)$$

$$predictors^{type1} =< congestion, mainroad, aod_{047}, aod_{055}, evi, ndvi, dem, temp, pop >$$

A tree ensemble model [10] developed in the form of K additive functions, which describe relations between AQI level with predictors,

$$aqi_{id} = \phi(predictors_{id}^{type1}) = \sum_{k=1}^{k} f_k(predictors_{id}^{type1}), f_k \in \mathcal{F}$$

To find an optimal model, a combination of hyperparameter tuning, feature selection, and a stratified 10-fold cross validation approach performed.

3.3 AQI level prediction using developed model and predictors data

Missing AQI level aqi' in grid cell i at day d with available predictors measurements (referred as type 2) predicted by using developed model,

$$aqi'_{id} = \phi(predictors_{id}^{type2})$$

3.4 AQI level inference where predictors data incomplete

Missing AQI level in grid cell where predictors data not complete for specific day aqi'' inferred by measuring association of grid cells AQI level values with AQI level located elsewhere and the association with available AQI values in neighboring grid cells. A universal Kriging Model developed with a smooth function of latitude and longitude and a random intercept for each cell.

$$aqi''_{id} = (\alpha + u_i) + (\beta + v_i)\ \overline{aqi'}_d + smooth(long, lat) + \epsilon_{id}$$

where $\overline{aqi}_d$ is the mean AQI level across bangkok on a day d. α, u_i, β_i and v_i are the fixed-random intercepts and fixed-random slopes, respectively. The smooth $long, lat$ is a spline fit to the latitude and longitude.

To measure the goodness of fit, Leave-one-out cross-validation performed by dropped 'all observations' of a sensors site from the dataset to become testing dataset for the stage-3 model. This process was repeated for each 11 sensors site and r^2 values were computed.

4 Results

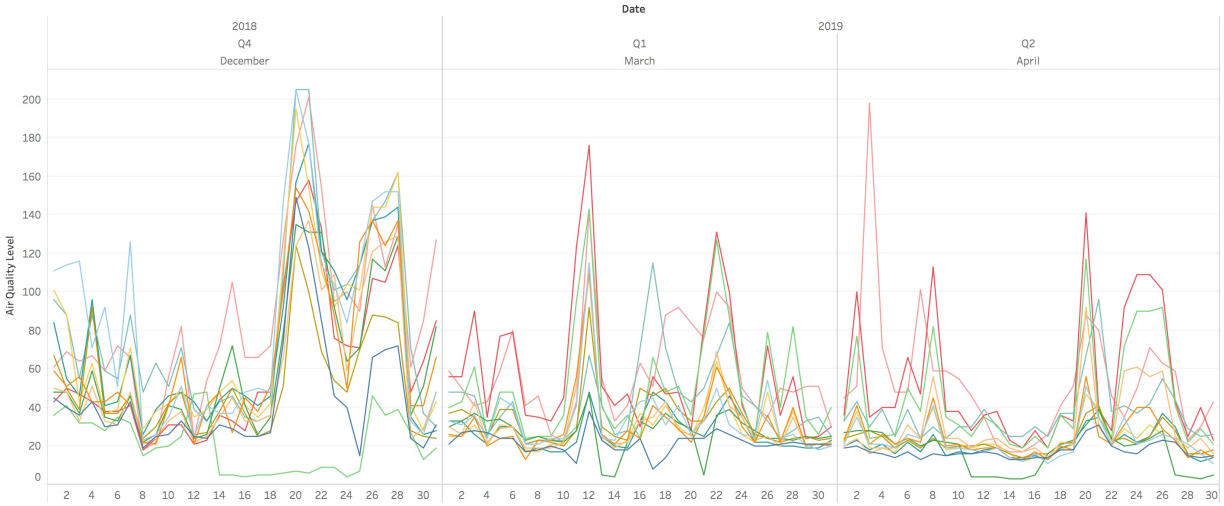
Bangkok Pollution Control Department provided three months daily aggregate AQI data from official sensors in Bangkok Metropolitan Area. Figure 3b shows ground sensors location along with daily measurements. As shown in Figure 2a there is one suspect of outlier sensors, the X12T site. It scores lowest correlation coefficient compared with others and therefore excluded from the next iteration.

The next step is to divide Bangkok into a grid of 1km x 1km grid follow by development of the tree ensemble model trained from ground truth and features from satellite and land use related data. From the best model, the location where there is no ground sensors but complete predictors could be inferred. The last step then extrapolation process to predict AQI level where there are missing predictors. Figure 2b shows performance of our model, the best model achieved 0.6 r^2 scores and also it can be seen from Figure 2c that the model can captures well the overall density distribution of actual AQI but may a bit overestimate it.

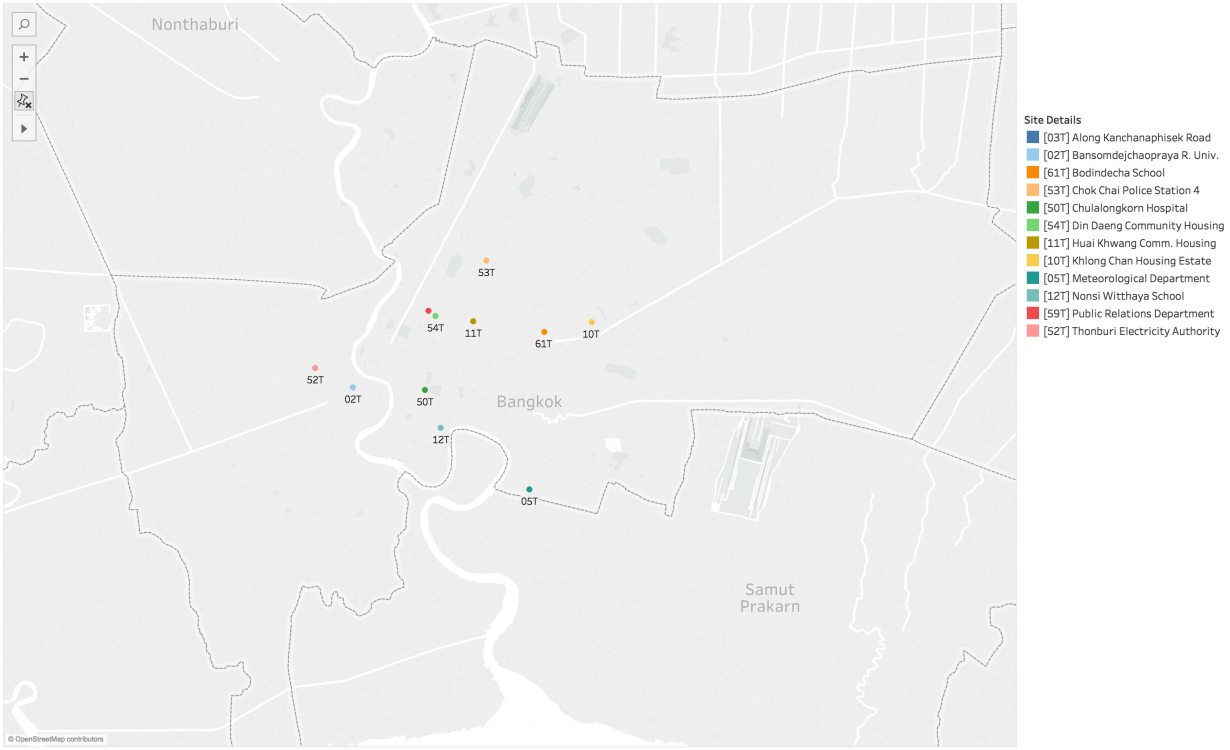
From the aforementioned processes, daily 1km x 1km AQI level from whole Bangkok could be inferred as seen in Figure 3 and 4. It can be observed there is difference between different time seasonality, the highest aqi level are observed during peak season ($\bar{aqi} = 74.5$), followed by normal ($\bar{aqi} = 54.9$) and low season ($\bar{aqi} = 41.8$).

Inferred AQI from our model are also able to gives complete situational awareness for example by calculating spatio and temporal aggregation by using districts boundary to create districts AQI statistics. As shown in Table I there are different pattern could be captured during different seasons. For the peak season the top-3 districts which have highest AQI level are Bangkok Noi ($\bar{aqi} = 89$), Nong Khaem ($\bar{aqi} = 89$), and Bangkok Yai ($\bar{aqi} = 88$). Different results are captured during normal season (Yan Nawa ($\bar{aqi} = 75$), Khlong San ($\bar{aqi} = 74$), Sathon ($\bar{aqi} = 71$)) and low season (Yan Nawa ($\bar{aqi} = 62$), Phra Nakhon ($\bar{aqi} = 61$), Ratchathewi ($\bar{aqi} = 61$)).

multirow



(a) Air Quality Level in Bangkok during December 2018, March-April 2019

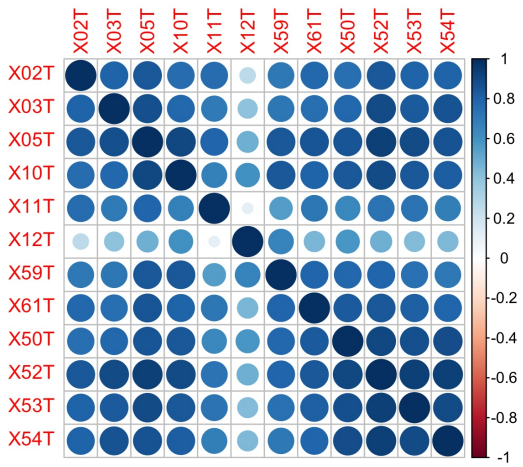


(b) Sensors Locations

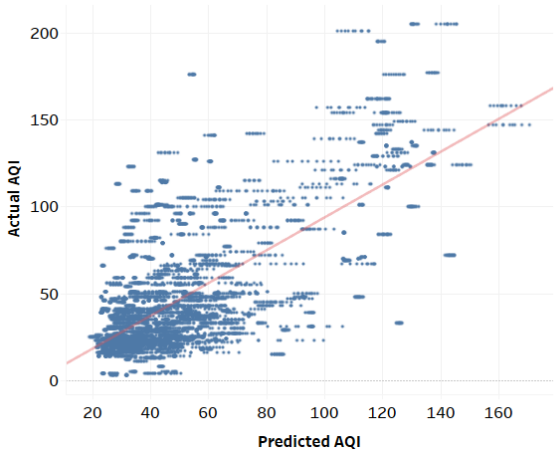
Fig. 1: Air Quality Data from Bangkok's Pollution Control Department

TABLE I: Districts AQI Statistics

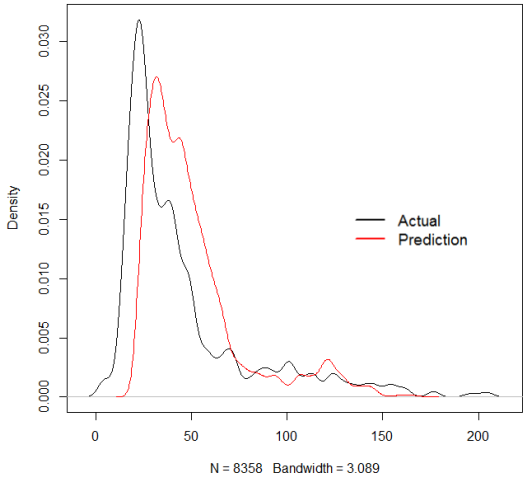
District	December 2018 (Peak Season)		March 2019 (Normal Season)		April 2019 (Low Season)	
	$\mu(mean)$	$\sigma(sd)$	$\mu(mean)$	$\sigma(sd)$	$\mu(mean)$	$\sigma(sd)$
Bang Bon	80	4	46	4	31	4
Bang Kapi	68	7	47	7	33	7
Bang Khae	82	5	48	8	35	6
Bang Khen	74	10	49	9	34	7
Bang Kho Laem	87	6	68	4	54	4
Bang Khun Thian	83	9	52	8	39	9
Bang Na	71	6	49	6	29	7
Bang Phlat	65	9	45	6	35	6
Bang Rak	67	4	67	3	58	2
Bang Sue	58	3	41	5	35	4
Bangkok Noi	89	8	66	6	53	10
Bangkok Yai	88	5	63	6	56	10
Bueng Kum	69	7	47	6	34	5
Chatuchak	63	8	39	7	32	7
Chom Thong	75	5	44	5	29	6
Din Daeng	83	6	63	6	50	8
Don Mueang	64	9	43	14	32	8
Dusit	77	9	65	7	57	8
Huai Khwang	62	7	47	4	32	4
Khan Na Yao	66	4	45	5	29	4
Khlong Sam Wa	74	12	56	8	41	7
Khlong San	75	7	74	5	59	3
Khlong Toei	84	11	70	3	57	4
Lak Si	67	8	47	7	35	6
Lat Krabang	77	7	54	8	34	6
Lat Phrao	68	8	47	5	33	5
Min Buri	74	8	51	7	37	7
Nong Chok	77	13	59	7	40	8
Nong Khaem	89	5	46	7	31	7
Pathum Wan	71	8	69	4	60	5
Phasi Charoen	69	6	41	8	31	7
Phaya Thai	67	6	46	8	34	8
Phra Khanong	73	12	52	11	34	8
Phra Nakhon	74	2	68	4	62	5
Pom Prap Sattru Phai	70	4	67	1	60	2
Prawet	74	8	54	7	34	5
Rat Burana	78	7	53	5	40	5
Ratchathewi	68	7	70	6	61	7
Sai Mai	72	10	52	7	33	6
Samphanthawong	73	5	65	2	57	1
Saphan Sung	73	8	50	7	34	7
Sathon	80	6	71	7	59	5
Suan Luang	69	10	52	7	34	6
Taling Chan	71	7	44	9	28	9
Thawi Watthana	80	7	44	6	32	5
Thonburi	85	7	66	9	57	6
Thung Khru	87	5	55	7	38	7
Vadhana	83	9	64	9	51	10
Wang Thonglang	69	6	49	6	34	6
Yan Nawa	87	12	75	3	62	6



(a) Correlation matrix for AQI ground sensor sites

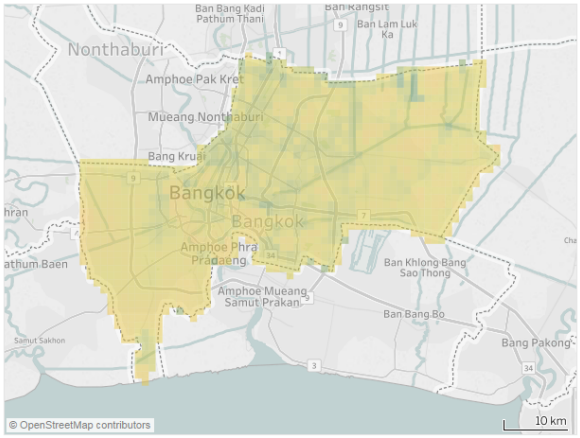


(b) Actual vs Predicted AQI level ($r^2 = 0.6$)

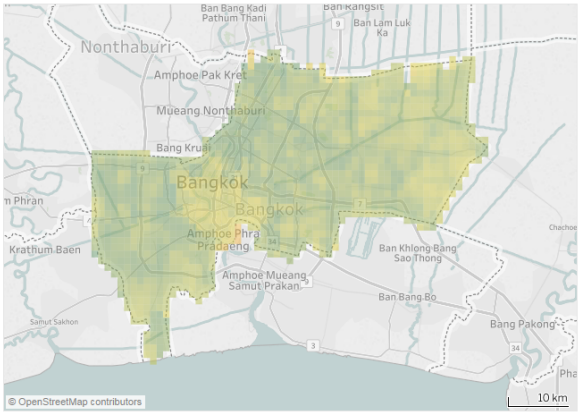


(c) Kernel Density plot of Actual vs Predicted AQI

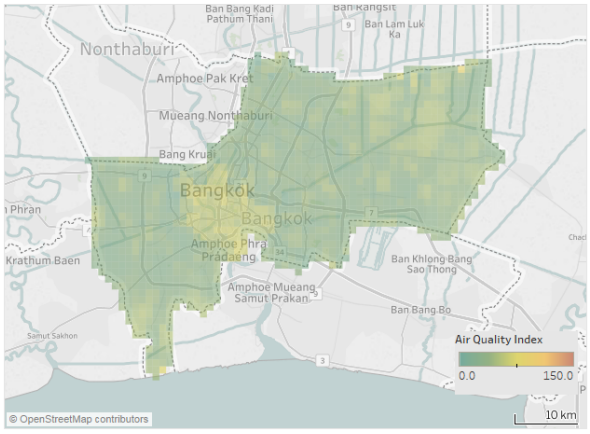
Fig. 2: Preprocessing and Model Performance



(a) Peak Season



(b) Normal Season



(c) Low Season

Fig. 3: Bangkok's air quality analysis results

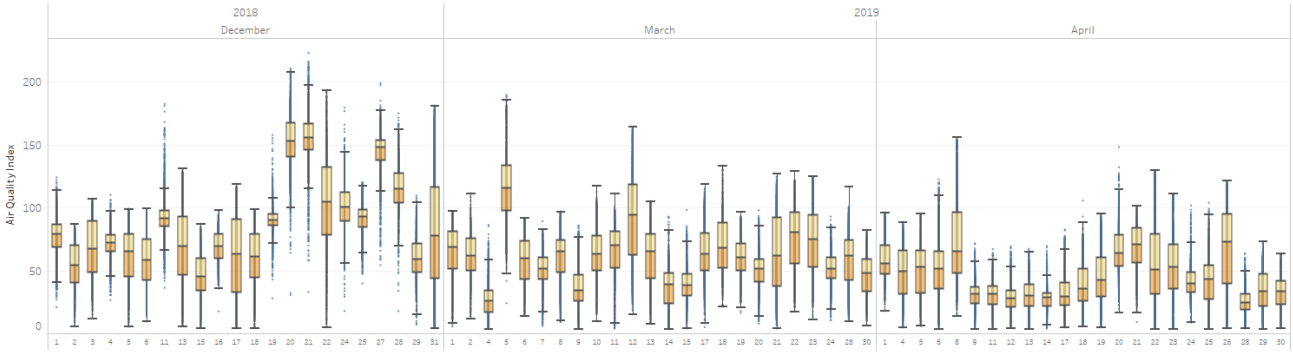


Fig. 4: Bangkok's air quality analysis results

Reflections and Further Works

1 The Potential of Alternative Data Sets

We explore the possibility of using ride-hailing data sets as alternative data sets to improve our understanding of travel patterns in Bangkok Metropolitan Region. The research shows the potential of ride-hailing data in inferring travel pattern, road-speed profile and traffic congestion nowcasting. Based on promising preliminary findings, we conducted comparison of traffic nowcasting from our model with Extended Bangkok Urban Model (eBUM) with mix results. We use traffic congestion nowcasting to measure population daily exposure to air pollution, better result is shown during peak traffic situation compare to normal traffic situation.

It is important to note that as proof-of-concept research, availability and access to different data sets is crucial for triangulation and means of verification. We used open data from sources such as Open Street Map and Worlpop when, in the beginning of the research, official statistics are not readily available or accessible. Office of Transport and Traffic Policy (OTP) and Policy Control Department of Thailand Government granted access to a set of official statistics on transportation, administration boundary and air quality information.

Immediate practical application will require further research to improve robustness of the model, in particular to include other data sources such as traffic counting, hourly air quality, land use, population and on-the-field observation. It is recommended to conduct further study in selected area in the city, based on government priority and primary data availability. Based on preliminary findings, it is recommended to continue improve road-speed profiling and traffic nowcasting model to compliment Extended Bangkok Urban Model (eBUM).

2 Partnership in Data Innovation

This research is a collaboration of GIZ Data Lab as project owner, GIZ Thailand as problem owner, Grab as data holder and Pulse Lab Jakarta as data analytics partner. The main beneficiary of this research is Government of Thailand and this is in line with Prime Minister of Thailand message to harness big data for decision making in 2018. The partnership managed to overcome the bureaucracy challenges that come for a non-traditional UN initiative to successfully partner with another non-traditional development sector initiative such as Data Lab and an emerging decacron such as Grab. We need to address Government of Thailand concern early in the implementation and able to do that with support from GIZ Thailand. This requires patience and trust from partners to ensure that there are shared value emerges for all partners.

3 Conclusions and next steps

While this work is preliminary, it does however show that combining alternate data sources with new techniques has the potential to augment the richness of insights that are available to developmental domains and in particular transportation planning and management.

3.1 Macroscopic traffic flow modelling

The analyses on leveraging ride hailing data in macroscopic traffic flow modelling explored two innovations. Firstly it incorporated a complementary data source that overcomes the timeliness issue of existing data used for traditional macroscopic traffic flow modelling which depends on infrequent surveys and censuses. And since it is transaction generated data (i.e. generated as a by-product of providing some service, in this case ride hailing), it is relatively costless to generate when compared to purpose built surveys and census that can be expensive and cumbersome to administer. Secondly the work explores the use of a radiation model and compares it to the gravity model, which is more commonly used in macroscopic traffic flow

modelling. Even at this preliminary stage, the radiation model shows promise, performing better than the gravity model. These results must be treated cautiously because in this preliminary work we find large deviations with actual traffic flows (inferred from the ride hailing data). But with additional work these newer models that also incorporate new data sources, can potentially enhance the eBUM traffic modelling currently utilized in Thailand. They could also potentially provide more accurate estimates of actual traffic flow for lower cost, with fewer data points but at a higher frequency and at higher spatio-temporal coverage. Irrespective of which model is utilized, they are both susceptible to localization errors which would require city specific calibration. Taking this preliminary work forward would mean refining the models further, integrating them into the existing eBum model, and conducting systematic validation of the outputs with ground truth data, ideally those from OD/ commuting surveys.

3.2 Road speed profiling

The advantage of using ride-hailing data for transportation specialists is its ability to provide a high frequency proxy estimate of traffic conditions and specifically overall traffic speeds. Given that speed information is captured continuously it also allows transportation specialists to understand the diurnal patterns of traffic and speed for different situations (e.g. weekday, weekend, special events, etc.). This enables the development of profiles for different types of activity that can each be updated in near-real time (subject to data access). Near real-time data like this can then potentially facilitate quicker feedback loops for transport agencies to trial and model effects of interventions (e.g. making roads one way). It can also inform the development of and testing of congestion charging policies.

Pneumatic tubes and induction loops which are most commonly used to capture vehicular speed on roads can be costly and therefore not feasible to deploy everywhere. It should be noted that ride hailing data doesnt directly give an estimate of vehicle counts or vehicle classifications which a combination of induction loops and pneumatic tubes can give. But when coupled with ground truth data and calibration, it would be possible to infer to some extent the volume of traffic on the roads from just the speed information from ride hailing services. Of course this comes with further caveats since ride hailing services usually cover only motorcycles and cars (in our preliminary work we only used data from cars), both of whose patterns of mobility and speed vary not just between themselves but also more importantly with public transportation such as buses which will have frequent stops. But as GPS on public transport also becomes more common, collectively with these different streams it would be possible to more improve the accuracy of the inference of volume of traffic. The deployment of CCTV can also improve the richness of the data that can be gleaned for transportation management and planning. Traditionally CCTV has not been very suitable for vehicle counts and vehicle classifications, but as costs of cameras continuously decrease and the state of the art of machine learning algorithms that can facilitate highly accurate vehicular counts and vehicular classification improves, it is possible that that deployment of these cheaper technologies coupled with access to data from ride hailing data can provide cheap continuous data of high frequency and high spatio-temporal coverage than what was possible before.

3.3 Traffic Congestion Nowcasting

By leveraging the ability to have a continuous flow of speed information, albeit with the caveats mentioned above, our preliminary research also looked at the possibility of developing a nowcasting traffic congestion model and compared its to outputs from the eBUM model. The results have been mixed with our model performing better on some roads and less so on others. This was to be expected given that we do not have rich ground truth data for validation purposes and also only leveraging one type of data (in this case vehicular speeds of cars from a ride hailing services). The model itself needs to be further improved so it is too soon to tell if our preliminary work represents a viable alternative and/or complement to the eBum model.

3.4 Inferring air pollution

Of the different use cases explored via this research, the last on inferring air quality at high spatio-temporal resolution directly tackles one of the

most important environmental and public health issues affecting several countries in South and South-East Asia. In the context of Bangkok, which currently only has 12 ground sensors, our preliminary work explored the use of AI techniques on data from multiple sources including, amongst others, satellite imagery and traffic congestion estimates to infer Air Quality at higher spatio-temporal resolution. Our works taking account of the Bangkok air quality conditions on various seasons, i.e. high, normal, and low seasons. However, data from just 12 sensors provides too little in terms of validation. Future work needs to have more ground truth validation data, which will also better help calibrate the model. But if this can be improved on and scaled, then it could provide mechanisms for a lower cost solution to having high spatio-temporal coverage.

4 Mainstreaming the use of alternate data sources

More work needs to be done to be able to mainstream this preliminary work into practise. This initial work is intended to inform policy enlightenment activities, to show the state of the art and the possibilities. The next significant step is to co-develop the design of rigorous research studies for each of these preliminary uses cases explored in this work. The design would have to provide for better ground truth data for validation and importantly to improve the preliminary work showcased here. As such having transportation specialists from Thailand involved both in the design as well as the implementation of this second phase would also provide for the cross-pollination of ideas and knowledge around the specific uses and limitation of the new data sources and techniques. range of rigorous research use cases to both provide better validation of this preliminary work but more importantly to improve it. This is all contingent also on having access to these alternate data sources. Even whilst this may be possible for the second phase of this work, sustainable longer term solutions to data access will need to be negotiated before this work can be mainstreamed.

References

[1] D. Milne and D. Watling, "Big data and understanding change in the context of planning transport systems," *Journal of Transport Geography*, vol. 76, pp. 235 – 244, 2019.

[2] R. Etminani-Ghasrodashti and S. Hamidi, "Individuals demand for ride-hailing services: Investigating the combined effects of attitudinal factors, land use, and travel attributes on demand for app-based taxis in tehran, iran," *Sustainability*, vol. 11, no. 20, 2019.

[3] "Ride-hailing - Asia, Statista market forecast," https://www.statista.com/outlook/368/101/ride-hailing/asia, accessed: 2019-12-15.

[4] Y. Yang, M. C. Gonzlez, A. Maritan, and A.-L. Barabsi, "Limits of predictability in commuting flows in the absence of data for calibration." *Nature*, vol. 4, no. 5662, 2014.

[5] M. Barthlemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1, pp. 1 – 101, 2011.

[6] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, Feb. 2012.

[7] A. Sunderrajan, J. Varadarajan, and K. Lye, "Road speed profiling for upfront travel time estimation," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. Los Alamitos, CA, USA: IEEE Computer Society, nov 2018, pp. 648–654.

[8] F. van Diggelen and P. Enge, "The worlds first gps mooc and worldwide laboratory using smartphones," in *Proceedings of the 28th International Technical Meeting of the Satellite Division of The Institute of Navigation*, 2015, pp. 361–369.

[9] P. Christidis and N. I. Rivas, "Measuring road congestion," *Institute for Prospective Technological Studies (IPTS), European Commission Joint Research Centre*, 2012.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[11] I. Kloog, P. Koutrakis, B. A. Coull, H. J. Lee, and J. Schwartz, "Assessing temporally and spatially resolved pm2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements," *Atmospheric Environment*, vol. 45, no. 35, pp. 6267 – 6275, 2011.

[12] I. Kloog, F. Nordio, B. A. Coull, and J. Schwartz, "Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal pm2.5 exposures in the mid-atlantic states," *Environmental Science & Technology*, vol. 46, no. 21, pp. 11 913–11 921, 2012.