Technical Report

The Sixth Research Dive on
Urban & Regional Development

May 2018

Australian Government

Executive Summary

Based on a report by the Organisation for Economic Co-operation and Development (OECD), roughly half of the world's population now lives throughout urban communities and that number is expected to increase in the coming decades. The challenges in urban area are manifold, and are especially intertwined in the areas of education, healthcare, access to clean water, *inter alia*. The emergence of new sources of data has created new approaches and opportunities to tackle many of these urban challenges, notably with access to information (and insights) that were not previously available from traditional datasets.

On 25-28 March 2018, twelve academics and researchers from across Indonesia, along with members from Pulse Lab Jakarta's team took part in the Lab's sixth edition of Research Dive for Development. The participants were divided into four research teams, each tasked with taking on a specific aspect of urban and regional development and assigned a dataset. Specifically, the tasks included: (1) Designing regional development policy, by analysing social events, news media data and its network based on GDELT (a global news media monitoring platform); (2) Assessing the accessibility to (emergency) health facilities in Sumatra, by analysing different datasets including transportation infrastructure distribution and health facility locations; (3) Monitoring water access for water supply infrastructure planning, by analysing several datasets including municipal waterworks customer distribution data; and (4) Inferring energy consumption towards urban development, by combining data on social media activity density and socio-economics statistics.

This technical report outlines the findings from the research sprint and is structured as follows:

1. The first paper describes the datasets that were assigned to the teams during the event.
2. The second paper explores the correlation between newsroom reporting and actual social events on the ground based on GDELT data, and discusses how the findings may be used to identify subject matters for regional cooperation and national development planning.
3. The third paper assesses the levels of exposure to two disasters of interest (flooding and haze) in Pekanbaru using multiple sources of data, and examines accessibility to health and potential evacuation facilities in order to propose ideal sites for shelter construction.
4. The fourth paper highlights observations from the results of two surveys from the Local Clean Water Company (PDAM) and the Indonesia Infrastructure Initiative (INDII), and proposes a solution to one of the major problems identified using supporting evidence.
5. The fifth paper investigates the correlation between the energy consumption by districts and the daily activities of Twitter users, and outlines a statistical model the team designed to infer the daily electricity consumption in Bali using daily Twitter activities.

Pulse Lab Jakarta is grateful for the cooperation of state-owned electricity company (PLN) Bali, Indonesia Australia Partnership for Infrastructure (KIAT), Institut Teknologi Bandung, Institut Teknologi Sepuluh Nopember, National Statistics Agency Nusa Tenggara Barat, Sekolah Tinggi Ilmu Statistik, Univesitas Andalas, Universitas Gadjah Mada, Universitas Indonesia, Universitas Jember, Universitas Multimedia Nusantara, Universitas Muhammadiyah Gorontalo, Universitas Padjajaran, Universitas Tanjungpura, Universitas Udayana, and Politeknik Ujung Pandang. Pulse Lab Jakarta is grateful with the support from DFAT Australia. Knowledge Sector Initiative (KSI) has committed to support the next research event.

Advisor Note

A Venue for Building Partnership and Networking

Researchers involved with urban and regional planning in Indonesia have been realizing how important it is to understand the role of modern "big-data" technology, especially in revealing more granular characteristics about a population. This kind of tool supports the efforts of city governments in developing people-oriented cities, whereby citizens are at the centre of sustainable development planning.

Being invited to participate in a 3-day research workshop at PLJ was an honor. It offered new perspectives and added useful knowledge through collaboration with a small group of researchers and data scientists, where together we explored and analysed various data sources such as traditional data media (newspapers) and modern digital media (social networking platforms). A partnership building and networking workshop in and of itself, PLJ creatively used Research Dive to bring to light relevant case studies that represent real world complex issues. I was thankful for having the opportunity to attend and take part in exploring possible solutions.

Dr. Ibnu Syabri
Advisor on Spatial Analysis

Dr. Ibnu is an Associate Professor in the Department of Regional and City Planning, SAPPK ITB. He received his bachelor's degree in Computer Science from the University of Kentucky and went on to complete his master's in Operation Research for General Engineering at the same university. In 2004, he was awarded a PhD in Transportation Economics from the University of Illinois in the United States.

An Innovative Event for Nurturing Research Ideas

Water supply infrastructure planning was the focus area for the group I mentored during the 3-days research dive event. Provided with data from a local clean water company (PDAM) and results from water socio-economic surveys conducted by IndII (Indonesia Infrastructure Initiatives) of Pontianak City, the participants in my group set out to address how the citizens, especially those among the urban poor, can efficiently access clean water. Given the fact that wetland area covers most of Pontianak, as well as that palm oil plantations grow fast in the mid-and-upper stream of the Kapuas river that crosses through city, there is a lack of access to clean water that calls into question the affordability and continuity of clean water supply in that region.

I was impressed with the passion and motivation of the team members, not only in diving into the available datasets, but also in seeking different data alternatives to enrich the discussion about the problem. One of their ideas was a 'communal water bank' that counts on the precipitation pattern. Coming up with such a promising concept in a short period time is commendable, and could be developed further in the design process and during its implementation in the city's management system. I was honored to be part of this event and thoroughly enjoyed the engaging discussions. I, for one, am looking forward to the further development of the communal waterbank. Thank you PLJ for introducing me to this innovative research event.

Dr. Hendricus Andy Simarmata
Advisor on Urban Planning

Dr. Hendricus has been a vice chairman of the Center for Urban and Regional Studies, Universitas Indonesia since 2015. In the last seven years, he has focused his work on integrating community resilience into urban planning. In 2011, he led a consultancy team commissioned by the Board of National Disaster Management to conduct risk assessment and disaster management planning. Mr. Simarmata earned his DPhil (PhD) in Development Studies from the University of Bonn, Germany in August 2016. He obtained his master's and bachelor's degree from the Department of Urban Studies, Universitas Indonesia in 2006 and the Department of Regional and City Planning, ITB in 2001, respectively.

Advisor Note

Digital Data to Better Understand Energy Consumption

Research Dive provides a new format for collaboration between academics and researchers from diverse backgrounds and with diverse competencies in Indonesia. I was pleased to be involved in this event as an advisor, focusing on energy consumption. The different professional backgrounds of the participants was never a barrier; they complemented each other and forged a genuine atmosphere for teamwork. The team I was assigned to had the task of inferring energy consumption towards urban development by combining social media activity density and socio-economic statistics.

It was really challenging for myself as an advisor and for the participants to explore non-traditional dataset of Big Data. Despite the limitations with the number of related tweets and PLN peak demand data from Bali Province in 2014, I was very impressed with how the team strategically formulated research questions, set up the methodology, crawled the dataset into meaningful information and then came up with brilliant ideas (and results) that were seemingly impossible at the beginning.

This event highlighted that with the presence of a digitised world today, understanding users' information from social networking platforms can enable policy makers to gain a better understanding about energy consumption and planning in Indonesia.

Dr. Lusi Susanti
Advisor for Urban Energy

Lusi Susanti is an Associate Professor in Industrial Engineering at the Faculty of Engineering, Andalas University. In 2004, she graduated from Toyohashi University of Technology, Japan with a master's degree and later completed her doctorate degree in 2008. She spent time took investigating various energy saving potential, in particular the reduction of building cooling load through naturally ventilated cavity roof. Recently, she has become more interested in observing urban energy consumption and energy saving potential through behavioural changes of building occupants. She has published several research articles in journals and conferences on this subject matter.

Research Dive

Advisors

Hendricus Andy Simarmata	Universitas Indonesia
Ibnu Syabri	Insitut Teknologi Bandung
Lusi Susanti	Universitas Andalas

Researchers

Group 1 – Designing regional development policy

Adiwan Fahlan Aritenang	Institut Teknologi Bandung
Guntur Budi Herwanto	Universitas Gadjah Mada
Novri Suhermi	Institut Teknologi Sepuluh Nopember
Imaduddin Amin	Pulse Lab Jakarta
Mellyana Frederika	Pulse Lab Jakarta

Group 2 – Assessing the accessibility to (emergency) health facilities in Sumatra

Febri Wicaksono	Sekolah Tinggi Ilmu Statistik
Putu Perdana Kusuma Wiguna	Universitas Udayana
Seng Hansun	Universitas Multimedia Nusantara
George Hodge	Pulse Lab Jakarta
Muhammad Rheza	Pulse Lab Jakarta

Group 3 – Monitoring clean water access for water supply infrastructure planning

Ahmad Komarulzaman	Universitas Padjadjaran
Ivan Taslim	Universitas Muhammadiyah Gorontalo
Muhammad Pramulya	Universitas Tanjungpura
Awan Diga Aristo	Pulse Lab Jakarta
Rajius Idzalika	Pulse Lab Jakarta

Group 4 – Inferring energy consumption towards urban development

Dharma Aryani	Politeknik Negeri Ujung Pandang
Dwi Martiana Wati	Universitas Jember
Wini Widiastuti	Badan Pusat Statistik - NTB Province
Muhammad Rizal Khaefi	Pulse Lab Jakarta
Pamungkas Jutta Prahara	Pulse Lab Jakarta

Table of Contents

Data Description for Research Dive Urban and Regional Development

Imaduddin Amin
Pulse Lab Jakarta
Jakarta, Indonesia
imaduddin.amin@un.or.id

Zakiya Pramestri
Pulse Lab Jakarta
Jakarta, Indonesia
zakiya.pramestri@un.or.id

ABSTRACT

The rapid pace of urbanisation has two sides: it is a fuel to economic and development growth; and it comes with the risks of social instability and related crises. Effective governance is key for managing these opportunities and challenges, which can be supported with insights from big data to understand the changing dynamics of a city. In other words, leveraging urban data to gain knowledge about the interactions among citizens, the interactions between citizens and city infrastructure, as well as the interactions between citizens and the environment is crucial for urban planning and governance.

To support the Government of Indonesia's efforts on better data utilisation for urban and regional development, Pulse Lab Jakarta organised a Research Dive for Development, where 12 researchers were invited to analyse multiple datasets. These datasets included: news media data, health facilities data, geospatial data of public health facilities and road network, water access data, electricity data and social media data. This paper describes the datasets that were used and is intended to contextualise the technical papers that follow. The datasets were provided by Pulse Lab Jakarta with the support of OpenStreetMap, Indonesia Australia Partnership for Infrastructure (KIAT), and PLN Bali.

1 INTRODUCTION

The world is currently facing rapid urbanisation. By 2050, 66 per cent of the world's population is expected to be living in urban areas[1]. Indonesia in particular is considered to be one of the top Asian countries with rapid urban growth, and too is expected to have more than half of its population living in cities by 2025[2]. These trends of rapid urbanisation are often perceived as a fuel to economic growth and development; however, through a different lens, rapid urbanisation is framed as a contributor to social instability, infrastructure crises, environmental hazards, among other unfavorable conditions. Thus, one of the key components for addressing some of these challenges and opportunities of urban development is effective governance.

With the proliferation of advanced digital technologies, effective governance also means making use of the vast amount of data available today. This provides more opportunities for city governments to understand the interaction among citizens, as well as between citizens and infrastructure, environment, and government structures in an urban context. Smart city and data-driven city are two urban development concepts that have been popularised, which are expected to assist with the monitoring, collection and analysis of various types of data in (near) real-time with the hopes of solving different urban problems and enhancing citizens' quality of life. There are numerous urban data sources these days, such as sensors, IoT, citizen-generated content, administrative data and customer transaction records from the private sector, yet the question remains: How can we leverage this abundance of data to support urban planning and governance?

The Government of Indonesia has been involved with promoting better data utilisation for urban and regional planning and development. Notably, the Ministry of National Development Planning has expressed interest in exploring new data sources to perform housing gap analysis, define metropolitan statistical areas, and identify proxies for the sustainable city index in order to enhance national programmes geared towards developing smart cities and green cities. Responding to the New Urban Agenda, the Ministry of Public of Works and Housing also has 35 areas for strategic regional development, with different focus within the development sector such as maritime, tourism, industry, logistics, agriculture, and services[3]. This plan involves an infrastructure integration component, which relies on insights from new data sources to examine the gap between existing conditions and the outlined goals in the plan.

Pulse Lab Jakarta organised its sixth Research Dive for Development, which focussed on areas of urban and regional development. The participants included 12 academics and researchers across diverse disciplines, namely urban planners, civil engineers, geography scientists, statisticians, computer scientists and economists. The objective was to extract insights to inform urban and regional planning and development, particularly on clean water access, energy consumption, health, and regional development.

Researchers were given access to news and social media data, health facilities data, geospatial data of public health facilities and road network, water consumption data, water survey data, and electricity data. The participants were divided into four groups and each assigned with a unique task: (a) to design a regional development policy, by exploring news from GDELT data; (b) to assess accessibility to (emergency) health facilities in Sumatra; (c) to monitor clean water access in Pontianak City in order to improve the water supply infrastructure; and (d) to infer energy consumption in Bali using aggregated social media data.

[1] http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html
[2] http://www.worldbank.org/en/news/feature/2016/06/14/indonesia-urban-story
[3] BPIW, "Profil Wilayah Pengembangan Strategis Indonesia", 14 March 2016, retrieved from http://bpiw.pu.go.id/

2 DATASETS

In this section, we briefly explain the four types of data that were made available under non-disclosure agreements to the participants for analysis during the Research Dive for Development.

2.1 News Media Data

2.1.1 GDELT Data. PLJ downloaded the news media data collected by The Global Database of Events, Language, and Tone (GDELT)[4]. Among the data compiled by GDELT, GDELT 1.0 Global Knowledge Graph or GKG was utilised, specifically the GKG Counts File (available from 1 April 2013).

GDELT GKG collects data globally from news article (archived from the previous day), and categorises each social events according to persons involved, related organizations, locations, quantities, and themes. These are then called a "nameset". Each nameset also includes a lists of news sources where the news in particular has been published. The GDELT GKG data is available in CSV format, each "nameset" having its own row. The data contains information as described in Table 1.

2.2 Health Facilities Data

2.2.1 Public Health Facilities Data in Riau. PLJ collected and cleaned the data on public health facilities in Riau (published by the Ministry of Health as of 2015). As shown in Table 2, the data includes information regarding each location, capacity for inpatients and outpatients, capacity to provide Basic Obstetric Neonatal Emergency Service, and its service coverage. The data was aggregated at the district level.

2.2.2 Hospital Location Data from OpenStreetMap. PLJ provided geospatial data of hospitals from OpenStreetMap, in shapefile format. The list of hospitals in OpenStreetMap was divided into 5 categories, namely: 'HEALTH_DENTIST', 'HEALTH_DOCTORS', 'HEALTH_HOSPITAL', 'HEALTH_PHARMACY', and 'HEALTH_VETERINARY'. The hospital locations were provided up to the sub-district level. Information on hospital location data is shown in Table 3.

2.2.3 Hospital Distance Travel Time Data from Google Street APIs. Referring to Google Street API, PLJ estimated the distance between the hospitals and the centre of the sub-district. In addition, the average time taken to reach the hospitals from the centre of the sub-district was estimated. Sample of data is shown in Table 4.

2.2.4 Road Network Data. PLJ provided geospatial data for road network in Riau in shapefile format. Figure 1 shows the road network map in Riau.

2.3 Water Access Data

2.3.1 Water Consumption Data. PLJ provided water consumption data in Pontianak, published by the regional water utility company PDAM Tirta Khatulistiwa[5]. The data is available for the timeframe April 2016 - March 2018. The data records water consumption of 200 consumers across Pontianak city.

The data contains information on the categories of consumers, water consumption, amount of charge, consumers' locations including details on latitude and longitude.

2.3.2 Water Survey Data. In partnership with Indonesia Australia Partnership for Infrastructure (KIAT), PLJ provided the survey data collected by Indonesia Infrastructure Initiative (INDII) project (established as of March 2017). The survey collected information on socioeconomic conditions and preferences on water system usage from 11,222 households across Indonesia. The questions sought to gather basic information about the household members (such as age, education level, and disability condition) and the sanitation condition of the household. The survey also covers each household's willingness to connect to the sewerage system and inquires about the amount of money each household would be willing to pay to get connected to a sewerage system.

2.4 Social Media and Electricity Data

2.4.1 Social Media Data. PLJ provided aggregated Twitter data, which records the volume of tweets and the number of unique user active per-30 minutes interval, covering Twitter activities in Bali province within 2014. The data is aggregated per-district.

2.4.2 Electricity Peak Demand Data. In partnership with PLN Bali, PLJ provided electricity peak demand data from all substations across Bali island, per-month for 2014. Information included in the data can be shown in Table 7.

3 DATA AND TASK MAPPING

The news media data was given to the first group, who explored the data to analyse the subject of interest from international perspective to shape regional policy. The second group used public health facilities data, geospatial data of public health facilities and road network in Riau, to assess the emergency health accessibility, particularly related to the haze and flood event. The third group had water access data, including water consumption data in Pontianak and water access survey result, to monitor the water access for better water supply infrastructure planning. Aggregated social media data was assigned to the fourth group, to infer the electricity consumption in Bali.

[4]https://www.gdeltproject.org/
[5]http://180.250.204.174/bacameter/index.php

Table 1: Example of GDELT data

Column Name	Type	Sample	Description
DATE	date (yyyymmdd)	20180320	Date of the news published
NUMARTS	int	40	Number of article/ source documents mentioning the count
COUNTTYPE	char	SOC_GENERALCRIME	Category of the count
NUMBER	int	5	Number of "OBJECTTYPE" being reported
OBJECTTYPE	char	Palestinians	Object of "NUMBER" refer to
GEO_TYPE	int	4	Category of geographic resolution. 1=COUNTRY, 2=USSTATE, 3=USCITY, 4=WORLDCITY, 5=WORLDSTATE
GEO_FULLNAME	char	Jerusalem, Israel (General), Israel	Geography name of location
GEO_COUNTRYCODE	char	IS	2-character FIPS10-4 country code
GEO_ADM1CODE	char	IS00	Country code followed by the 2- character FIPS10-4 administrative division 1
GEO_LAT	numeric	317.667	Centroid latitude of the location
GEO_LONG	numeric	352.333	Centroid longitude of the location
GEO_FEATUREID	signed int	-797092	GNS or GNIS FeatureID for this location
CAMEOEVENTIDS	char	740198791,740200871,...	GlobalEventIDs
SOURCES	semicolon-delimited list	english.wafa.ps...	List of sources publishing related articles
SOURCEURLS	semicolon-delimited list	http://english.wafa.ps/page...	List of source URLs publishing related articles

Table 2: Data of Public Health Facility in Riau

Column Name	Type	Sample	Description
KODE	char (4)	1401	4 digits of District ID
KAB/KOTA	text	KUANTAN SINGINGI	District name
RAWAT INAP	int	11	Number of inpatient capacity
NON RAWAT INAP	int	12	Number of outpatient capacity
JUMLAH	int	23	Number of total capacity
KODE PUSKESMAS	text	P1401010102	ID of public health facility
NAMA PUSKESMAS	text	Lubuk Jambi	Name of public health facility
LINTANG	double	-0.685197	Latitude
BUJUR	double	101.4528891	Longitude
PONED/ NON	text	PONED	Types of public health facility, whether it is capable to provide Basic Obstetric Neonatal Emergency Service
KEMAMPUAN PENYELENGGARAAN	text	Rawat inap	Capability of public health facility to serve inpatient and outpatient, or outpatient only
LUAS WILAYAH	double	397.00	Area of region
WILAYAH KERJA DESA	int	21	Number of villages service coverage
JUMLAH PENDUDUK	int	20717	Number of population within service coverage

Table 3: Hospital Location Data from OpenStreetMap

Column Name	Type	Sample	Description
NAME	text	RSIA Norfa Husada Bangkinang	Hospital Name
CATEGORY_ID	text	HEALTH_ HOSPITAL	There are five main categories such as 'HEALTH_DENTIST', 'HEALTH_DOCTORS', 'HEALTH_HOSPITAL', 'HEALTH_PHARMACY', 'HEALTH_VETERINARY'
ID_PROV	char	14	2 digits of Province ID
ID_KAB	char	1406	4 digits of District ID
ID_KEC	char	1406050	7 digits of Sub-District ID
PROV	text	Riau	Province Name
KAB	text	Kampar	District Name
KEC	text	Bangkinang	Sub-District Name
LAT	double	0.345223	Latitude
LONG	double	101.027209	Longitude

Table 4: Hospital Distance & Travel Time Data from Google Street APIs

Column Name	Type	Sample	Description
NAME	text	Rumah Sakit Ibu dan Anak Labuah Baru	Hospital Name
CATEGORY_ID	text	HEALTH_HOSPITAL	There are five main categories such as, 'HEALTH_DENTIST', 'HEALTH_DOCTORS', 'HEALTH_HOSPITAL', 'HEALTH_PHARMACY','HEALTH_VETERINARY'
ID_PROV	char	14	2 digits of Province ID
ID_KAB	char	1471	4 digits of District ID
ID_KEC	char	1471011	7 digits of Sub-District ID
PROV	text	Riau	Province Name
KAB	text	KOTA PEKANBARU	District Name
KEC	text	PAYUNG SEKAKI	Sub-District Name
LAT	double	0.517031999	Latitude
LONG	double	101.4264	Longitude
ORIGIN	string	MARPOYAN DAMAI	Sub-District Name
DURATION	int	1361	Duration in seconds
DISTANCE	int	9485	Distance in meters

Table 5: Pontianak Water Consumption Data (From PDAM Tirta Khatulistiwa)

Column Name	Type	Sample	Description
idpelanggan	int	1010024	Id number of user
ukuran	string	Diameter 0.50	Size of PDAM pipe
iddiameter	int	1	Id of size mentioned above
kodegol	string	2A2	Code of customer category
golongan	string	RT Semi Permanen	Customer category
stanskrg	int	2048	Water meter shown now (m^3)
stanlalu	int	2023	Water meter shown in the previous month (m^3)
pakaiskrg	int	25	Water consumption (m^3)
totalrekening	int	94000	Amount of water charge (IDR)
idkelainan	int	15	Id of abnormalities condition
kelainan	string	ST - Stan Tempel [MMNU]	Abnormalities condition
idrayon	int	28	Id of sub-district area (rayon)
rayon	string	M-124	Sub-district area (rayon)
alamat	string	JL.RAJAWALI-GG.RAJAWALI NO.51	Address of customer
latitude	double	-0.0228848954570614	Latitude
longitude	double	109.332	Longitude

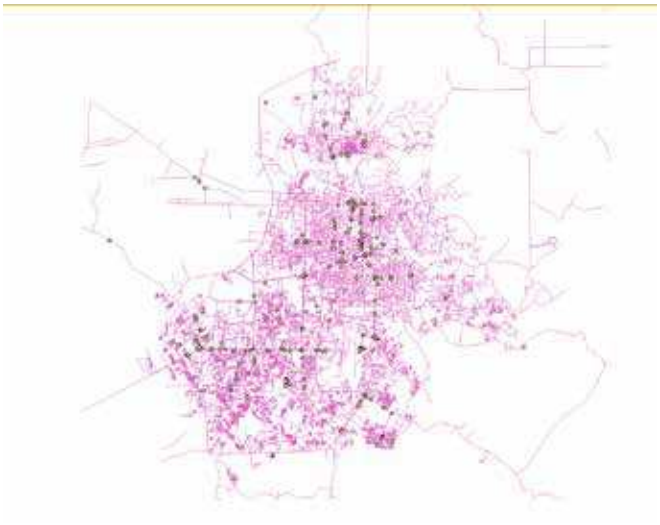


Figure 1: Geospatial data of road network in Riau

Table 6: Aggregated Twitter Data

Column Name	Type	Sample	Description
DATE	Date (dd/mm/yy)	01/01/14	Date of record
HOUR	int	8	Hour of record
MINUTE	int	0	Minute of record (0 or 30)
PROV	int	51	Code of province
KAB	int	5102	Code of region
KEC	int	5102011	Code of district
PROV_NAME	string	Bali	Name of province
KAB_NAME	string	TABANAN	Name of region
KEC_NAME	string	SELEMADEG TIMUR	Name of district
SOURCE	string	Others	Source of tweet (web/ others)
UNIQUE_USER	int	2	Number of unique user active during the 30-minutes record
NUMBER_OF_TWEETS	int	2	Number of tweets posted during the 30-minutes record

Table 7: Electricity Peak Demand Data per month (From PLN Bali)

Column Name	Type	Sample	Description
GARDU INDUK (SUBSTATION)	text	Gianyar_Trafo_1	Name of substation
JUMLAH PELANGGAN	int	166568	Number of consumers covered by substation
DAYA (MVA)	int	60	Electricity power distributed from substation
NOMINAL (A)	int	1732	Electrical current distributed from substation
BEBAN PUNCAK (A) - SIANG	int	757	Electricity peak demand at afternoon within a month
BEBAN PUNCAK (A) - MALAM	int	966	Electricity peak demand at night within a month

Shaping Regional Development Policy Through News Coverage: Potential and Limitation

Adiwan Aritenang
Insitut Teknologi Bandung
Bandung, Indonesia
a.aritenang@gmail.com

Guntur Budi Herwanto
Universitas Gadjah Mada
Yogyakarta, Indonesia
gunturbudi@ugm.ac.id

Novri Suhermi
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
novri@statistika.its.ac.id

Imaduddin Amin
Pulse Lab Jakarta
Jakarta, Indonesia
imaduddin.amin@un.or.id

Mellyana Frederika
Pulse Lab Jakarta
Jakarta, Indonesia
mellyana.frederika@un.or.id

ABSTRACT

For public policy to address a problem timely and correctly, ability to respond to an actual, perceived or anticipated problem is critical. Big data holds tremendeous potential in providing information for policy analyst that is more timely, accurate and detailed. This paper examines the use of big data to shape public policy, to look on the potential of worldwide news capture in The Global Data on Event Location and Tone (GDELT) project inform policy making in Indonesia context. GDELT data sets reveal topic of interests from the following neighbourhood countries: Australia, Singapore and Malaysia on selected Indonesian big islands namely Sumatera, Kalimantan, Jawa, Bali and Papua. We concluded that big data from news play a role in shaping foreign perceptions towards specific Indonesian regions which could be responded and anticipated by policy analyst in that regions.

KEYWORDS

GDELT Data, Public Policy, Indonesia

1 INTRODUCTION

For public policy to address a problem timely and correctly, ability to respond to an actual, perceived or anticipated problem is critical. Policy analyst have been using large, high-dimensional data sets as evidence to policy making. In the advance of technology, there are new sources of digital data known as big data available for policy analysis.

Shintler and Kulkarni argue that big data holds tremendous potential for public policy analysis, new resource for helping to inform different points in the policy analysis process, from problem conceptualization to ongoing evaluation of existing policies and even empowering and engaging citizens and stakeholders in the process. Big data can be useful in producing information that is more timely, accurate and detailed than that gleaned from more traditional sources of data [3].

This paper examines the use of big data to shape public policy, to look on the potential of worldwide news capture in GDELT project inform policy making. The Global Data on Event Location and Tone [1] database contains nearly a quarter of billions geocoded records on global events going back to 1979 and collects 100,0000 news events every day. The news from more than 100 languages has been translated into English and, uses natural-language processing turned the news into data points. It is one of the largest open-access spatio-temporal datasets with total archives span more than 215 years, GDELT provides a wealth and unprecedented amount of information on global societal system and behavior [2].

We ask the following question "How can we use worldwide news inform and advise us to shape our development policies". This is done by (i) understanding the real-world events from GDELT data, (ii) understanding the connection between sources and events and its connection to the real-world, and (iii) examining how a specific regional development policy in Indonesia can be formulated based on the world news.

2 RESEARCH METHODOLOGY

2.1 Data

We downloaded GDELT 1.0 Global Knowledge Graph, specifically the GKG Counts File that are available since April 1st, 2013. GDELT GKG collects data globally from news article from the previous day, and pairs a set of person names, organization names, locations, counts, and themes that then being called a *nameset*. The data is available in CSV format, with each unique *nameset* per row.

We focused on events happened in Indonesia, it is about 242,500 events out of 42 millions event available from GDELT 1.0. We limit the research to the potential of GDELT data to policy making in Indonesia context.

2.2 Methodology

2.2.1 Descriptive Analysis. To understand GDELT data reflection to real-world event, we conducted descriptive anslysis by visualizing the simple count of news by date as seen in Figure 1. The graph shows top news on Indonesia are on natural disaster such as earthquake, volcano eruption and flood, and news on terrorism. The datasets provide near-realtime insights into what is happening in different places in Indonesia.

2.2.2 Location Analysis. The next step is understanding the connection between sources and events and its connection to the real-world. We infer sources of newsroom by identifying server location and domain name such as .au that indicates an Australian newsroom and .my that indicates a Malaysian newsroom. We notice the limitation of this approach. First, there are newsrooms with

[1] www.gdelt.project.org

server located outside their country such as United States of America. Second, there are newsroom that use .com or .co instead of country-related domain name.

The data sets consists of event's location up to city level. However, our analysis reveal list of news without specific location and incorrect location. Based on this finding, we analyse news content to identify location of the events based on selected big islands in Indonesia namely Sumatera, Kalimantan, Java, Bali and Papua. These are big islands that came at the top of the news list. Hence, other big islands such as Sulawesi and islands in East and West of Nusa Tenggara, Maluku and North Maluku Province are not selected.

2.2.3 Topic Analysis. GDELT data sets has given 44 categories such as kill, arrest, wound, protest and more. This is a broad categorization and focused on crisis and violent events. We created different category by using topic analysis called Latent Dirichlet Allocation (LDA) to identify news topic and category automatically. We identified 9 different topic models with three different news corpus. In order to create the corpus, the news article is transformed into a corpus with a bag-of-words form. The Corpus became an input for LDA topic modeling. Based on this, we preprocessed the news through the following steps: a) elimination stopword b) dictionary formation c) forming document matrix and d) creation of corpus with bag-of-words form. We enhanced the above-mentioned topics by manually adding relevant keywords according to researchers' knowledge (keyword spotting) [1].

2.2.4 Analyse GDELT data sets with international visitors datasets. As the last step, we combined the information from GDELT with Indonesia Statistics (Badan Pusat Statistik) data sets on number of international visitor per country of origin in the year of 2014 and 2015.

3 RESULTS AND DISCUSSION

Each neighboring country has different topic of interests towards specific Indonesian island. In the past five years, Australia main interest is the Bali Nine case. It is the name given to group of nine Australian convicted for attempting to smuggle heroin out of Indonesia in April 2005 and the execution of two convicts happened in April 2015. Singapore is the only country of origin that reported haze events. We found reports on forest-related issues such as plantation, orang utan, elephant from newsroom located in Malaysia and Singapore. We noted that forest-related issues are appeared only on news about Sumatera and Kalimantan. Hence, there is specific characteristic of an island that attracts certain topic. For Papua, topic around human right is appeared.

The newsroom located in Australia has significant coverage on Bali while newsroom located in Singapore and Malaysia shows a large interest in Sumatera and Kalimantan. The bordering regions also exhibit the area of cross-culture and political dimensions. The following examples showcase result of topic modeling on understanding main interests of neighboring country: a) Malaysia main topics are on Rohingya refugee, Jakarta governor election, tourism, and also the flood disaster, b) Singapore main topics are on forest fire, mining, tourism and gay issues and c) Australia main topics are on volcano eruption.

We compare this to Indonesian-based news on the three countries. We found the following coverage: a) on Malaysia: palm oil and batik b) on Singapore: tax amnesty, and c) on Australia: refugee and tourism.

We use international visitors statistics to see the correlation between the number of international visitors with the bad news reported by the country of origin of the international visitors. One of the results of the analysis can be seen in Figure 6. During bad news, the number of international visitors from Singapore and Australia to Indonesia is reduced.

4 CONCLUSION

GDELT data sets reveal real-world events in Indonesia with some limitations. GDELT data set suffers from temporal bias and captures only events with news-value. Category with the biggest number of nameset is 'kill' and keywords such as poverty is not in the top list. Create a different category by automatically create new topics combined with domain expert input can be valuable to capture information on specific issues such as tourism, politics and human rights issues. A further study on the regional development policy can provide better context to the automatic topic.

GDELT data is rich resources on global news that provide potential new sources for national and local analysis. Issues that are important for Australia, Singapore and Malaysia newsroom on Indonesian islands can be different with national and local perspective and this can be used to understand the different perspective and to provide appropriate policy towards foreign perspective. A further study is required to understand the different perspective between national and foreign news and how that affect regional development in Indonesia context.

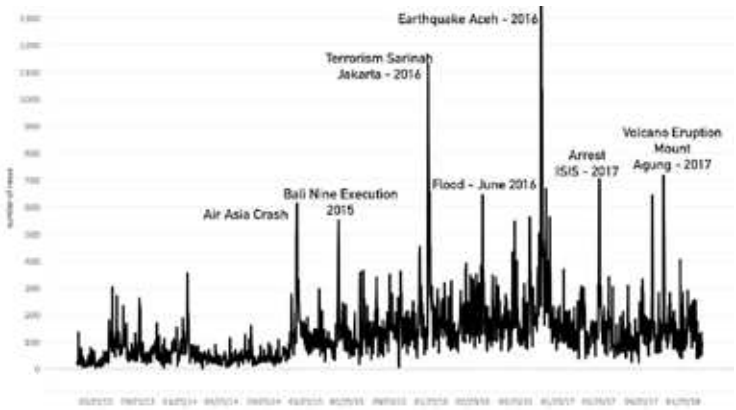
However, GDELT database suffers from temporal bias. To avoid implementation of inappropriate or inequitable policies, it is important to understand the extent and nature of bias in the data, and if possible correct for it.

REFERENCES

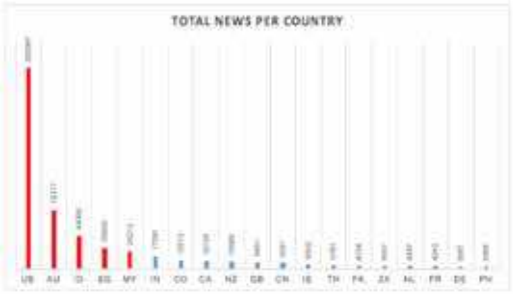
[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* (2003). http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[2] Kalev Leetaru and Philip A. Schrodt. 2013. (2013). http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

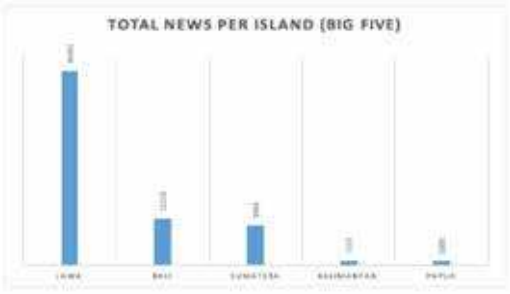
[3] Laurie A. Schintler and Rajendra Kulkarni. 2014. Big Data for Policy Analysis: The Good, The Bad, and The Ugly. *Review of Policy Research* 31, 4 (jul 2014), 343–348. https://doi.org/10.1111/ropr.12079



(a) News Count by Date



(b) Total News per Country



(c) Total News per Island (Top Five)

(d) Indonesia News Distribution across Indonesia on "Kill"

(e) Indonesia News Distribution across Indonesia on "Poverty"

Figure 1: News Count and Distribution from GDELT platform

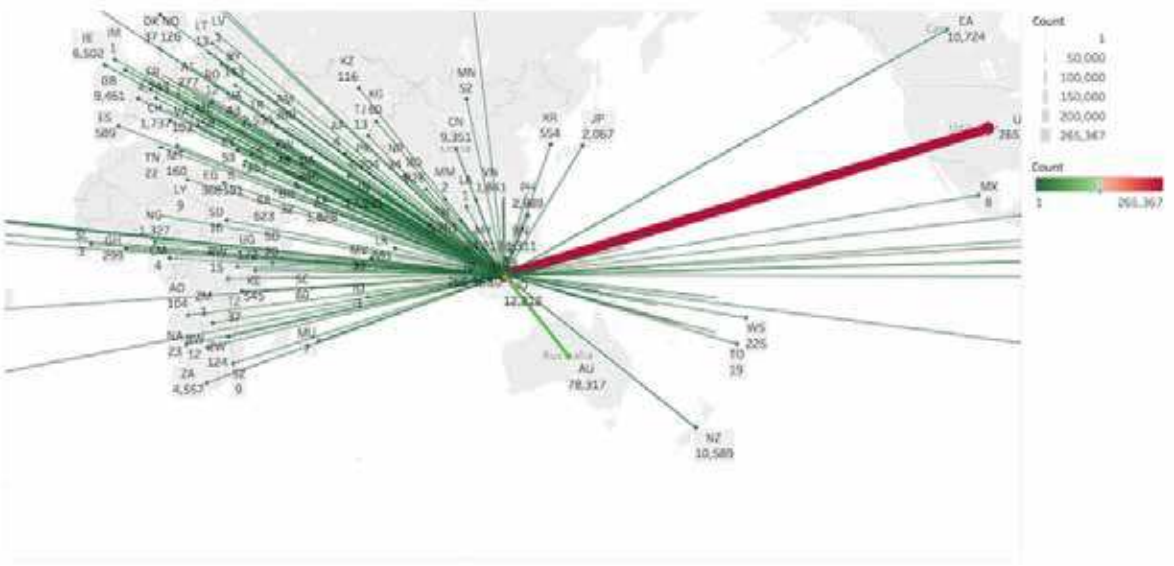


Figure 2: Indonesia News From Around the World

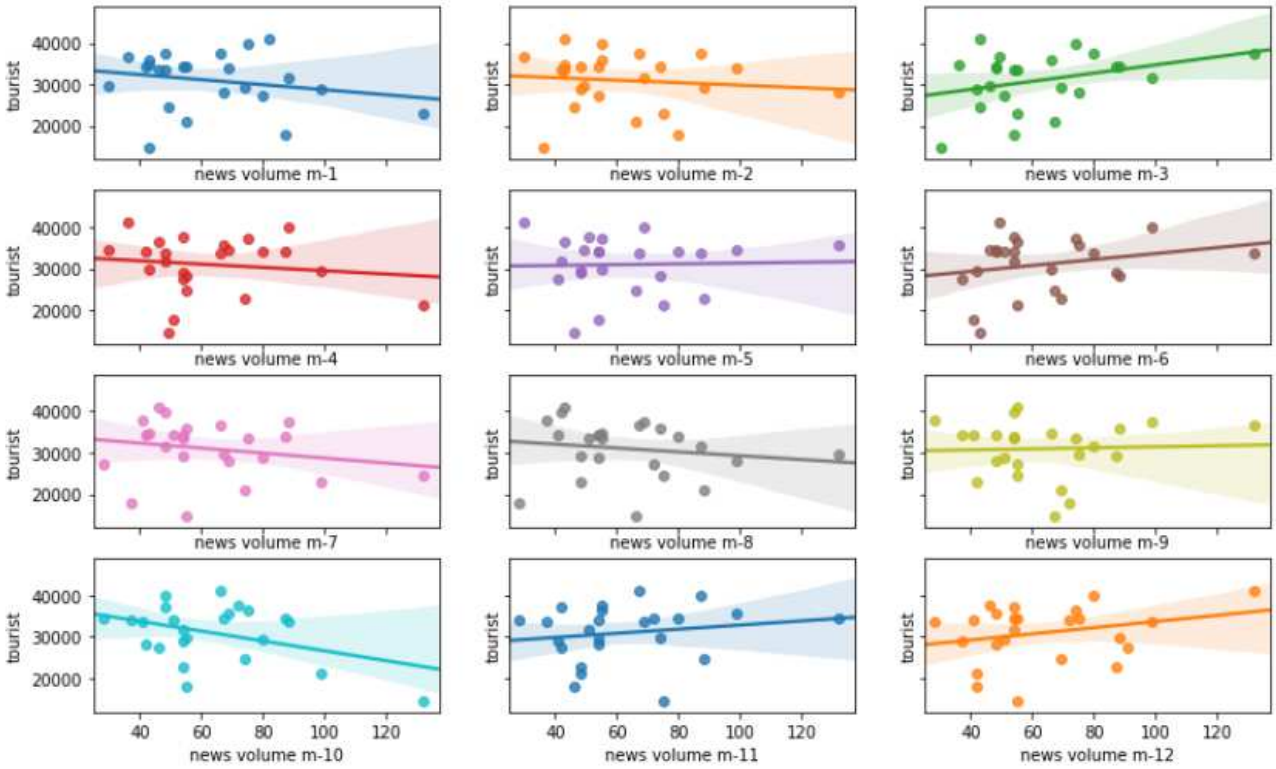


Figure 3: Bad News and Tourist Correlation

Spatial Accessibility of Health Facilities in Relation to Disaster Hazards in Sumatra : Case Study in Riau Province

Putu Perdana Kusuma Wiguna
Center for Spatial Data Infrastructure Development, Universitas Udayana
Denpasar, Indonesia
putu.perdana@gmail.com

Febri Wicaksono
Sekolah Tinggi Ilmu Statistik (STIS)
Jakarta, Indonesia
febri@stis.ac.id

Seng Hansun
Program Studi Informatika
Universitas Multimedia Nusantara
Jakarta, Indonesia
hansun@umn.ac.id

Muhammad Rheza
Pulse Lab Jakarta
Jakarta, Indonesia
muhammad.rheza@un.or.id

George Hodge
Pulse Lab Jakarta
Jakarta, Indonesia
george.hodge@un.or.id

ABSTRACT

Health sector is a crucial factor for the development of a country. A common problem in establishing and developing the health sector is the optimal distribution of health facilities that affect their accessibility. In this study, we focused on the analysis of spatial accessibility of health facilities in Riau Province. Moreover, we relate the health facilities location common disaster hazards in Riau Province (i.e. forest wildfire and flood). From the study, we successfully identify and classify accessibility of health facilities while taking into account the disaster hazards.

KEYWORDS

Health facilities, disaster hazards, spatial accessibility, Riau Province

1 INTRODUCTION

The health sector is a very crucial factor for the development of a country. In part because good health is linking with longevity, happiness and to the productivity of the economy. It is no coincidence than that good health and well-being is one of 17 goals of United Nations Sustainable Development Goals (SDG) [1]. In contrast with Millennium Development Goals (MDGs), the health SDG has a broader scope, which inferred from its goal, "Ensure healthy lives and promote well-being for all at all ages". Figure 1 depicts the importance of Health sector in the SDG Era.

According to a World Health Organization (WHO) report, the 2016 World Health Statistics, many countries remain far away from universal health coverage [2]. It implies that there is still much work to be done by governments and other stakeholders to increase national health coverage. One important task related to this matter is the analysis of health facilities location and by extension their spatial accessibility.

Access to healthcare facilities can have multiple definitions, but for the purposes of this research we segregate it into two parts, i.e. namely potential for healthcare delivery and the realized delivery of care, as suggested by Jamtsho and Corner[4] and Guagliardo [3]. The potential for healthcare delivery refers to the potential need for health care services of the entire population in the healthcare servicing region, while the realized delivery of care refers to the actual utilization of healthcare services by the population in need [4]. Accessibility itself is spatial in nature and in many literature we find the term 'spatial accessibility' (SA), which can be defined as the study of spatial components of healthcare accessibility. In this research, we will focus on the SA of the healthcare system in Sumatra, specifically in Riau province.

Figure 1: SDG 3-Health

Riau is a province in Indonesia with Pekanbaru as its capital city. With a total area of 96409.54 km^2, it has 10 regencies (Kabupaten) and 2 cities (Kota) [2]. Riau province is located on Sumatra island with geographic location ranging from 01005'00"S to 02025'00"N and from 100000'00'E to 105005'00"E [2]. Like other regions on Sumatra island, Riau has different topographies. In the east, it has lowlands; in the west, it has highlands; and in the center part, it mainly consist of undulating terrain. Figure 2 shows a map of Riau province taken from a Badan Perencanaan Pembangunan Daerah (BPPD) report [2]. The biggest Regency is Indragiri Hilir with a total area of 14161.74 km^2 and the smallest Regency is Pekanbaru City covering 683.46 km^2. The Cities and Regencies in Riau Province are shown in Table 1.

There are many reports of disaster hazards in Riau province. We will focus on two of the most common disaster hazards, namely

[1] http://www.un.org/sustainabledevelopment/health/
[2] http://www.un.org/en/sections/issues-depth/health/

Figure 2: Map of Regencies in Riau province

Table 1: Regencies and Cities in Riau Province

Regency/City	Area(km^2)	Area %
Rokan Hilir	9881.92	10.25
Rokan Hulu	7968.89	8.27
Kuantan Singingi	5686.84	5.90
Indragiri Hulu	8547.57	8.87
Indragiri Hilir	14161.74	14.69
Pelalawan	13937.67	14.69
Siak	8393.92	8.71
Kampar	11711.09	12.15
Bengkalis	9119.24	9.46
Kepulauan Meranti	3867.32	4.01
Pekanbaru	683.46	0.71
Dumai	2449.88	2.54
Total	**96409.5**	**100**

flood and wildfire. Almost every year, Riau experiences flood, especially in its lowland areas. In 2017, floods in Riau affected at least 2,467 households and more than 10,000 individuals [3]. As for wildfires, Riau has known as the province with the highest fire hotspot density in Sumatra based on data from 2006 to 2015 [1].

Geographic Information Systems (GIS) are an emerging technology for spatial analysis in healthcare studies [7]. It has also become a potential tool for assessing the geographic distribution of health services [5], since it can offer accurate measurements of spatial accessibility. There are some reliable GIS software, both proprietary and open source, such as ArcGIS (proprietary) and QGIS (free). There is also a specific health-related GIS software, namely Access

[3]http://mediaindonesia.com/read/detail/94992-korban-banjir-riau-mencapai-10-391-jiwa

Table 2: Data Sources

No	Data	Source
1	Riau's regency (Kabupaten) boundaries (figure 2)	Riau Spatial Planning Map
2	Riau's road network	Riau Spatial Planning Map
3	Distribution of Health facilities (Hospital and Puskesmas)	OpenStreetMap and Google Earth
3	Hotspot locations	Haze Gazer (Pulse Lab Jakarta)
5	Digital Elevation Model	USGS

Mod[4][6], which is developed and supported by the World Health Organization (WHO).

Based on description above, the main problem we are trying to address in this study is to analyze the spatial accessibility of health facilities and potential shelters in relation to forest fires and flood hazards in Riau province.

2 MATERIALS AND METHODOLOGY

2.1 Materials and Data

As explained in Introduction section, this study utilized open source data and software. We use QGIS software, which is a free and open source Geographic Information System software, which downloaded from https://www.qgis.org/en/site/. By using it, we can create, edit, visualize, analyze, and publish any geospatial information on different platforms [5]. By using it, we can create, edit, visualize, analyze, and publish any geospatial information on different platforms. The latest testing version of QGIS is 3.0.1, however for this study we used the latest stable version, i.e. 2.18.18 LTR.

The latest testing version of QGIS is 3.0.1, however for this study we used the latest stable version, i.e. 2.18.18 LTR. Data on this research is using opensource data. The opensource data are spatial data that collected from various sources, as shown in Table 2.

The data are explained as follows:

(1) **Road Network Data**. The road network data can be seen in Figure 3. Road in Riau Province is divided into five types: Toll Road, Primary Arterial Road, Primary Collector Road, Secondary Collector Road and Primary Local Road. The total road length is 4267.78 km with the longest road being a Primary Arterial Road with total length of 1158.33 km or 27.14% of the total. Table 3 shows the length of each road type in Riau Province.
(2) **Health Facilities Distribution**. Health facilities distribution identified using Google Earth and Openstreet Map data. Distribution of health facilities are shown in Figure 4.
(3) **Hotspot Location**. Hotspot locations from forest and peatland fires in Riau Province are taken from data covering 2014, with one dot representing one fire. Figure 3c shows the hotspot map.
(4) **Digital Elevation Model (DEM)**. The United States Geological Survey's (USGS) Digital elevation models (DEMs) are

[4]https://www.accessmod.org/
[5]https://www.qgis.org/en/site/

Table 3: Distance from Main Road

No	Distance	Score
1	<100 m	5
2	100 m - 250 m	4
3	250 m - 500 m	3
4	500 m - 1000 m	2
5	>1000 m	1

Table 4: Distance from City/Regency Center

No	Distance	Score
1	<5km	4
2	5km - 10km	3
3	10km - 15km	2
4	>15km	1

arrays of regularly spaced elevation values referenced horizontally either to a Universal Transverse Mercator (UTM) projection or to a geographic coordinate system. The grid cells are spaced at regular intervals along south to north profiles that ordered from west to east. Figure 6 shows the DEM of Riau Province. DEM is used as data input for flood hazard model.

2.2 Research Steps

In this section, we explain the research methodology for this study briefly, as follows.

(1) **Task and research question formulation.** On the first day of Research Dive (RD) 6, a specific task to be solved by Pulse Lab Jakarta (PLJ). Then, we brainstormed some basic ideas and research questions that can be related to the given task.
(2) **Literature and methods review.** Next we conduct a literature review and study related methods to solve the problem we have. We also identify some data types and collections that will be used in this research.
(3) **Material and data collection.** In this phase, we collected all the data we needed for the research as had been explained in section 2.1. The data then mapped and analyzed in the next phase.
(4) **Data mapping and analysis.** Using all the data we have, we map it to address research questions. We used QGIS as supporting tool in analyzing the data. The analysis involves DEM (Digital Elevation Model) analysis, buffer and heat map analysis to produce classes and scores of accessibility for health facilities. Health facilities in this research is restricted to only Hospitals and Public Health Centers (Puskesmas). Classes of accessibility of health facilities are based on distance from a main road, distance from City/ Regency centers and distance from hazard prone areas. Health facilities that are located in hazard prone area are not included in the analysis as it is assumed they would be equally affected by the hazards and thus less capable of providing assistance. Table 4 and Table 5 shows the classes and score for each class. Highest total score used as parameter of accessibility to health facilities. Five classes of accessibility are related to the total score, namely, very easy, easy, moderate, difficult and very difficult.
(5) **Findings and reporting.** In this phase, all research findings related to the research question are reported in a technical report paper. Figure 7 describes the research methodology in more detail.

3 RESULT AND DISCSUSSION

Figure 8 displays the areas prone to wildfire (hotspot) in Riau Province (the last recorded data was in 2014). From the figure, we can see that hotspot-prones areas spread widely covering the northern and central parts of Riau province. The most affected Regencies are Rokan Hilir, Bengkalis, Dumai, Siak, Kepulauan Meranti, Pelalawan, and Indragiri Hilir. The least affected Regencies are Rokan Hulu, Kota Pekanbaru, Indragiri Hulu, and Kampar. Only the Regency of Kuantan Singingi is unaffected by hotspots.

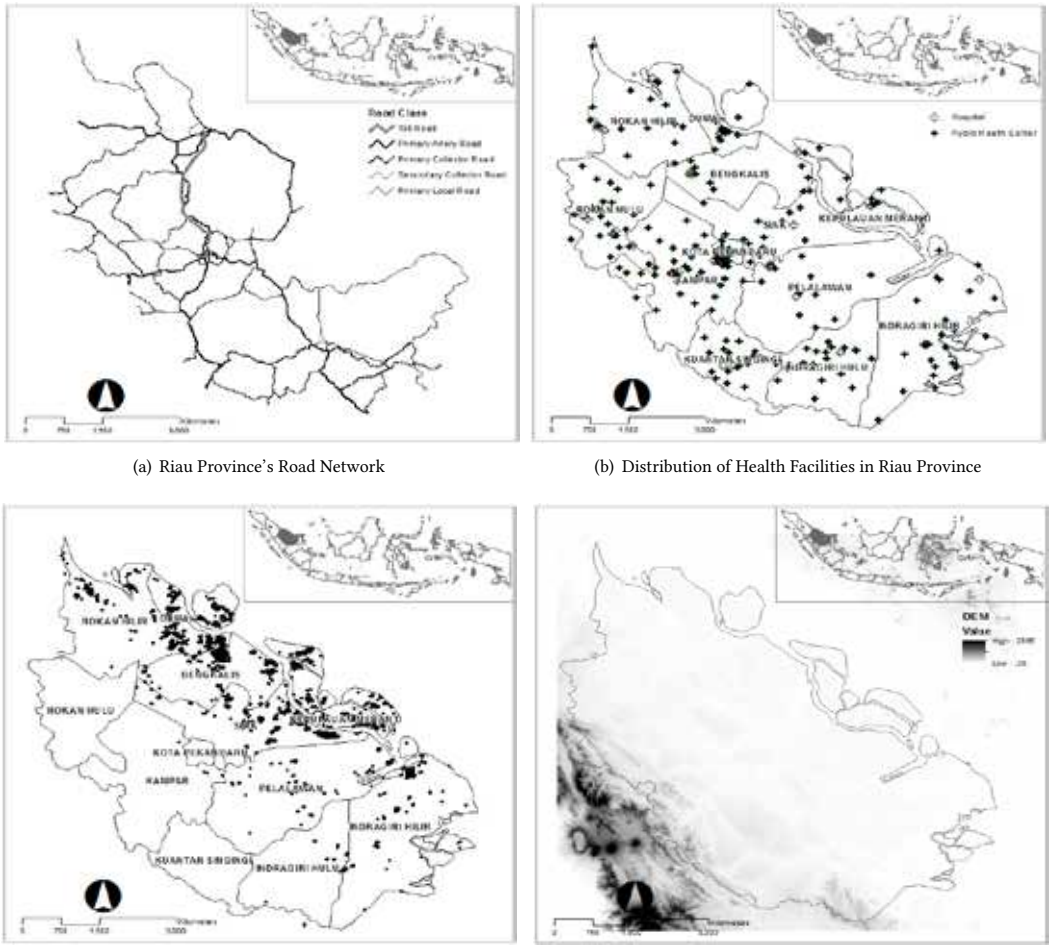
The flood events are also widely spread across the northern part of Riau Province. The most affected Regencies are Rokan Hilir, Dumai, Kepulauan Meranti, and Indragiri Hilir. The least affected Regencies are Pelalawan, Kota Pekanbaru, Kampar, Bengkalis, Siak, and Indragiri Hulu. Only two Regencies are unaffected by flood hazard, namely Rokan Hulu and Kuantan Singingi. Figure 9 shows the flood hazard distribution in Riau province.

From Figure 8 and 9, it's clear that disaster hazards are widely spread across the northern part of Riau Province. Therefore, analysis of access to health facilities should focus on the ones that are located outside of the hazard area. Figure 10 shows the distance of health facilities (hospitals and health centers/ Puskesmas) from from their Regency/ City centers. Fifty eight health facilities are located less than five kilometers from their Regency/ City centers, meanwhile twenty six health facilities are located from five kilometers to ten kilometers from city center. Figure 11 shows the number of health facilities within several distances from their Regency/ City centers.

Figure 12 shows the health facilities distance from a main road in Riau province. Only seven health facilities are located less than 100 meters from a main road, mostly are located more than 1000 meters from a main road. Only fifteen health facilities are located within 100 m to 500 meters from a main road. Figure 13 shows the number of health facilities with distance categories from main road.

Using both the health facilities' location and distance from main roads, we then classify the health facilities's spatial accessibility into five categories, namely, very easy to access, easy to access, moderate to access, difficult to access, and very difficult to access. Figure 14 shows the accessibility to health facilities in Riau Province regarding disaster hazards.

Figure 15 shows the number of health facilities and their accessibility related to disaster hazards. Three health facilities have very easy access and eight health facilities classified as easy access related to disaster hazards. This is very small proportion of the total. Eighty health facilities classified as difficult to very difficult accessibility to disaster hazards.



(a) Riau Province's Road Network

(b) Distribution of Health Facilities in Riau Province

(c) Riau Province's Hotspot Distribution in 2014

(d) Digital Elevation Model Map

Figure 3: Geospatial Data of Riau

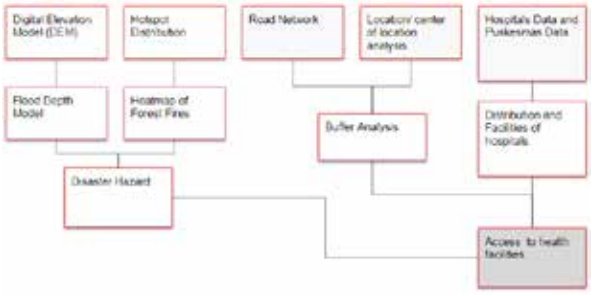


Figure 4: Research Methodology Flow Diagram

4 CONCLUSION

Only three health facilities have very easy access and eight health facilities are classified as easy access related to disaster hazards. This also means that the majority of the health facilities are difficult to access with 80 health facilities classified as difficult to very difficult accessibility related to disaster hazards. From this study, there are also some recommendations for the Government and related stakeholders:

- The needs to improve road access to health facilities;
- The needs for more equal distributions of health facilities and infrastructures that can lead to better accessibility of health facilities;
- The importance of media to spread awareness of disaster risks and their relationship to the location of available health facilities

5 ACKNOWLEDGMENTS

The research team wishes to acknowledge and thank Pulse Lab Jakarta for the support in terms of data access, facilities, and research dive logistics.

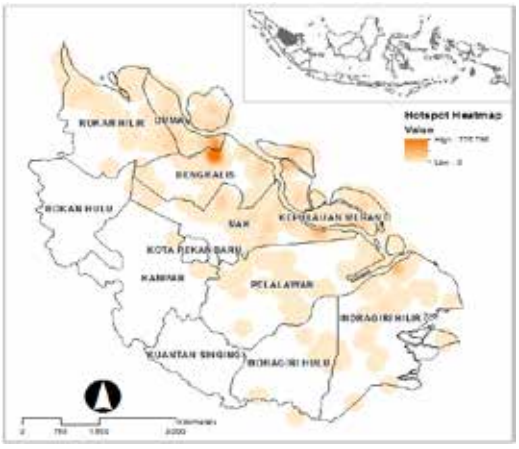


Figure 5: Hotspot Analysis Results

Figure 6: Flood Distribution

REFERENCES

[1] I. Albar, I.N.S. Jaya, B.H. Saharjo, and B. Kuncahyo. 2016. Spatio-temporal typology of land and forest fire in Sumatra. *Indonesian Journal of Electrical Engineering and Computer Science* (2016).

[2] Badan Perencanaan Pembangunan Daerah (BPPD). [n. d.]. RPJMD Provinsi Riau Tahun 2014-2019,Pemerintah Provinsi Riau. ([n. d.]). arXiv:https://www.bappenas.go.id/

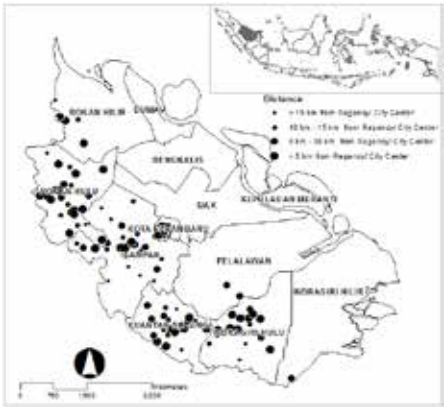
[3] M.F. Guagliardo. 2004. Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics* (2004).

[4] S. Jamtsho and R. J. Corner. 2014. Evaluation of spatial accessibility to primary healthcare using GIS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (2014).

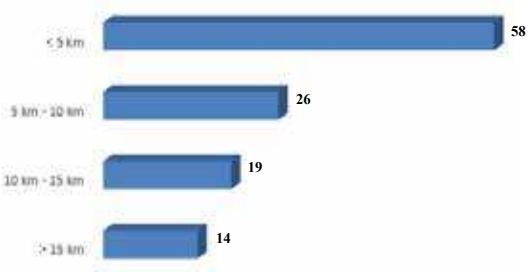
[5] S. Mansour. 2016. Spatial analysis of public health facilities in Riyadh Governorate, Saudi Arabia: a GIS-based study to assess geographic variations of service provision and accessibility. *Geo-spatial Information Science* (2016).

[6] N. Ray and S. Ebener. 2008. AccessMod 3.0: computing geographic coverage and accessibility to health care services using anisotropic movement of patients. *Geo-spatial Information Science* (2008).

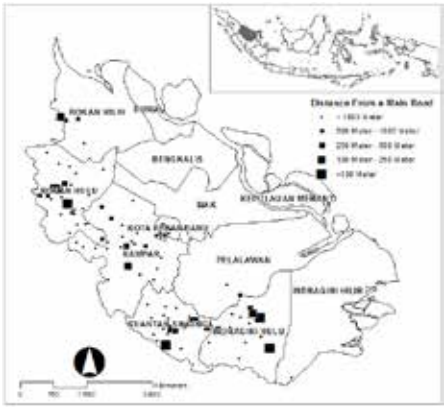
[7] L.Y. Wong, B.H. Heng, J.T.S. Cheah, and C.B. Tan. 2012. Using spatial accessibility to identify polyclinic service gaps and volume of under-served population in Singapore using Geographic Information System. *International Journal of Health Planning and Management* (2012).



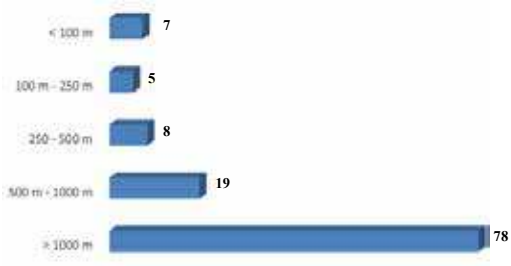
(a) Health Facilities' Distance from Their Regency Centers



(b) Health Facilities' Distance from their Regency Centers

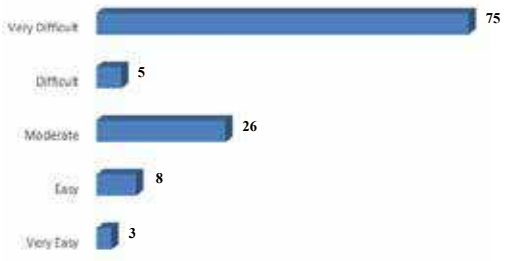


(c) Health Facilities' Distance from Main Roads



(d) Health Facilities' Distance from their Regency Centers

(e) Accessibility of Health Facilities in Riau Province



(f) Accessibility of Health Facilities in Riau Province

Figure 7: Health facility distance and accessibility

Monitoring Clean Water Access for Water Supply Infrastructure Planning in Pontianak City

Muhammad Pramulya
Universitas Tanjungpura
Pontianak, Indonesia
muhammad.pramulya@gmail.com

Ahmad Komarulzaman
Universitas Padjajaran
Bandung, Indonesia
ahmad.komarulzaman@unpad.ac.id

Ivan Taslim
Universitas Muhammadiyah Gorontalo
Gorontalo, Indonesia
ivantaslim@umgo.ac.id

Awan Diga Aristo
Pulse Lab Jakarta
Jakarta, Indonesia
awan.aristo@un.or.id

Rajius Idzalika
Pulse Lab Jakarta
Jakarta, Indonesia
rajius.idzalika@un.or.id

ABSTRACT

The service system of clean water network in Pontianak provided by the state water company (PDAM Tirta Jaya) in average has high coverage across the city. Yet, some parts of the community are still left behind and require attention. This study aims to reveal which part of the community receive the least service, their current coping mechanism. It also aims to identify the potential solutions that PDAM or local community might be able to consider at an affordable cost. We utilise various datasets related to water, PDAM customer database, socio-economic background and spatial related information. Based on those, we identify that majority of low income households in sub-district Pontianak Utara do not have formal access to water provision by PDAM. Our finding on the behavior of non-piped water households confirms that these are the most vulnerable community and we propose an alternative solution to expand the water supply network by building a water bank using rainwater and runoff especially accessible in the northern area of Pontianak City

KEYWORDS

Water supply company, water access, clean water, water bank

1 INTRODUCTION

Clean water accessibility is undoubtedly a part of basic human right. Water is as important as air and food to maintain the basic function of human body especially for health. However, in many underdeveloped and developing countries, clean water deprivation remains a critical issue. The technology on water distillation can address the clean water access, especially to prepare drinkable water from sea water, dew, and swamp.

However, the invention of water distillation is expensive. Given that water infrastructure is a major problem for communities, such as in rural or remote areas, the solution for tackling this challenge is to consider local capacity and norms, as well as accommodate the demand side from people.

In this paper, we will focus on Pontianak as subject of the study, located in West Borneo. Geographically, the city is located in the intersection between Kapuas and Landak river, two big rivers in Kalimantan [9]. The Kapuas River of Borneo is Indonesia's largest river system. Measuring 1,143 kilometres, it is also the world's longest island river [3],[12] characterised by a complex geomorphology and a intricate network of waterways and hydrological links with surrounding bogs and wetlands [1],[4], [5]. Bogs and wetlands are increasingly important land resources for livelihood, economic development, and terrestrial carbon storage [1],[4], [5]. The city's future agricultural development depends on this environmentally fragile peatland because of the dominance (58% and 16% area, respectively) [1].

The city relatively flat topographic contours with a ground level between 0.1 to 1.5 meters above sea level [2]. Pontianak City has a tropical climate that is divided into two parts: the rainy season and dry season[3,11]. In normal conditions, the dry season occurs in May to July while for the rainy season occurs in September to December. The average air temperature reaches 28-32 Celcius with air humidity ranges between 86% - 92% and the duration of solar irradiance 34-78% [8]. The amount of rainfall ranges from 3,000-4,000 mm per year with an average wind speed reaches five to six knots per hour [7].

Administrative wise, it consists of six sub-districts, namely Pontianak Kota, Pontianak Selatan, Pontianak Utara, Pontianak Tenggara, Pontianak Timur and Pontianak Barat. According to National Statistics Agency [11], sub-district Pontianak Utara (and to a less extent Pontianak Timur) is lagged behind in terms of general infrastructure and income level. Hence, it is more likely that residents in Pontianak Utara have a lower level of clean water access [8].

In order to come up with a local solution for the neediest places, we need to answer the following research questions:

(1) How does the spatial distribution of PDAM users explain water access inequality, particularly between Pontianak Utara and other sub-districts?
(2) How is the existing non-piped households' behaviour pattern in accessing clean water?
(3) How can we improve clean water access for low income households in Pontianak Utara, given the current nature, existing infrastructure development and the demand side of non-piped households?

[1] http://bappeda.pontianakkota.go.id/statis-16-profilfisikdasarkotapontianak.html
[2] http://bappeda.pontianakkota.go.id/statis-16-profilfisikdasarkotapontianak.html

2 DATA AND METHODS

2.1 Data

This study utilizes two main data sources that complements each other.

(1) Pontianak PDAM Tirta Khatulistiwa Data that provides the information water consumption and geo location of the piped water consumers in Pontianak City April 2016 - March 2018.
(2) INDII water dataset to get the socio-economic characteristics and water use pattern of households without piped water in Indonesia in 2017.

The secondary data includes information such as population data, data of regional facilities and infrastructure, topography data, and climatology data. All secondary data were obtained at the institution in the study area such as Central Statistics Agencies (BPS) Pontianak, Tirta PDAM Equator, Public Works Department and other related institutions.

2.2 Methods

We employ two equations to obtain a fair estimate of water demand and supply, that can further be the basis of our planning for water provision.

2.2.1 Water needs (Wn) Assessment in Pontianak Utara. Assuming that our work is focused in Pontianak Utara, the calculation of water requirement in Pontianak Utara (liter/ person/day) is done based on existing standard or guidance such as Indonesian National Standart (SNI). Clean water requirement according to SNI 03-7065-2005 about the standard of clean water requirement in a region is 120 liter/ person/day in Indonesia. So the formula to get the amount of water needs in the area of North Pontianak is:

$$Wn = TP \quad x \quad 120 \tag{1}$$

Wn :Water needs (liter/day)
TP :Total of Populations (2016)
120 :Indonesian National Standart (SNI) (liter/person/day)

2.2.2 Rainwater Harvesting (RH). Rainwater Harvesting (RH) is a method used in utilisation in an area during the annual rainy season. For that reason, it is necessary to know the average annual rainfall estimation in the North Pontianak area, so that it can be planned the amount and the area of catch required in making "water bank" in order to fulfill the water requirement.

$$RH = Fc \quad x \quad Ra \tag{2}$$

RH : Rainfall Harvesting (m3)
Fc : Field of catchment (m2)
Ra : Rainfall average (mm/day)

3 RESULT AND DISCUSSION

3.1 How does the spatial distribution of PDAM users explain water access inequality?

First of all, we would like to find out the scale of water infrastructure inequality in Pontianak City by analysing state water company customer database, and if this is related to the previously identified spatial variation by BPS data.

The coverage of piped water in Pontianak City is relatively high, but the services are not distributed evenly over the city. PDAM Tirta Khatulistiwa, state water supply company in Pontianak City, serves about 89% of households in the southern part of the Kapuas river, compare to 66% in the northern part (Fig. 2).

This is in line with income inequality found by BPS and our projection (Fig.1) that Pontianak Utara is relatively less improved in key development dimensions compared to other sub-districts.

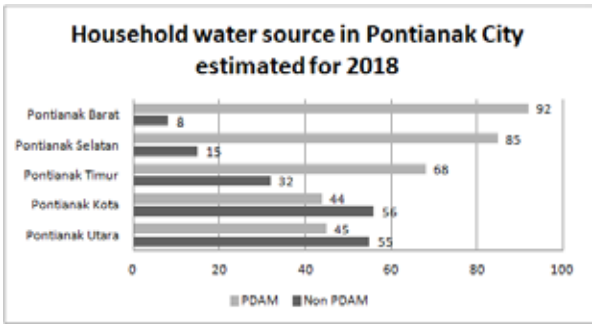


Figure 1: Estimated Household Water Source in Pontianak City for 2018. Source: own elaboration from The Integrated Database for the Social Protection Program

Figure 2: Water Company Service Coverage in Pontianak City

3.2 How is the existing non-piped households' behaviour pattern in accessing clean water?

After getting the evidence of the spatial inequality, we move into the second research question that focusing on behavior of households with non-piped water. This step gives a support to adjust our approach for water provision intervention.

We find that households without piped water connection tend to have multiple sources of water depending on their income (Fig.3).

Poor households mostly rely on single water source while richer households have more resources to access at least two water sources. Poor households mostly relies on dug well and others water sources (direct extraction from rivers or rain water harvesting) for drinking and bathing purposes and to a lesser extend buy refillable bottled water and vended water for drinking purpose. Meanwhile, the richer households choose better quality and relatively more expensive water sources such as branded and refillable bottled water for drinking and pump well for bathing.

In terms of water expenditure, there is a positive correlation between income and spending for drinking water (Fig. 5a). The bathing curve, however, decreases for the highest income bracket (Fig. 5b).

These findings indicate several aspects. First, poor households buy a smaller quantity of water for drinking but with a higher price than the richer households (Fig. 4a). Second, a larger amount of water for bathing is not affordable for poor households. Hence, we expect that poor households rely on dug well, rain water harvesting and water from open sources that are relatively inexpensive but of low quality (Fig. 4b). Third, given the relatively low average water expenditure for bathing purposes among rich households, this can be explained by the fact that most of them have pumped well that is free of charge (Fig. 4c). This indication is confirmed by the fact that poor households need to allocate 8.8% of their income to fulfill their basic water needs, far more than the commonly acceptable affordability rate at 4% (Fig. 4d). This indicate that non-piped water is relatively not affordable for the poor.

Connecting the first two research questions, we identify that poor households in Pontianak Utara are the most vulnerable, and at the same time most of them rely on alternative ways yet more pricey to obtain water for drinking and other households purposes. This leads to the third research question.

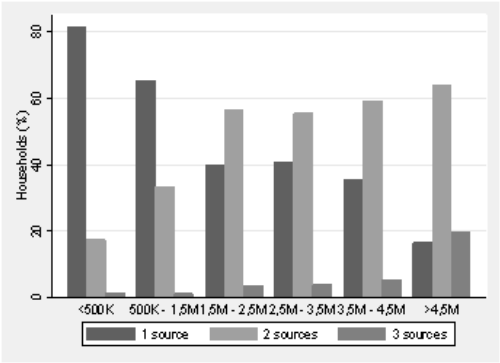


Figure 3: Number of main sources of water for drinking, cooking and bathing by households income level. Source: Authors calculation from Water INDII Survey 2017

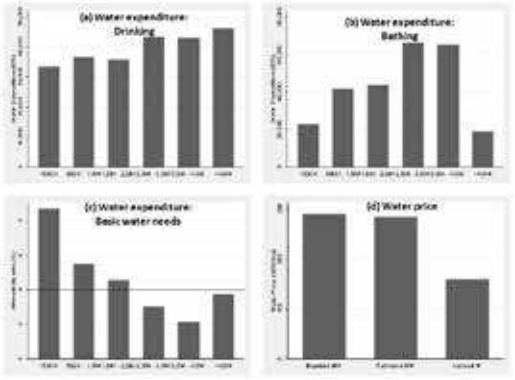


Figure 4: Average household expenditure on water for (a) drinking, (b) bathing, (c) basic water needs by income level and (d) average price of water by Water INDII Survey 2017

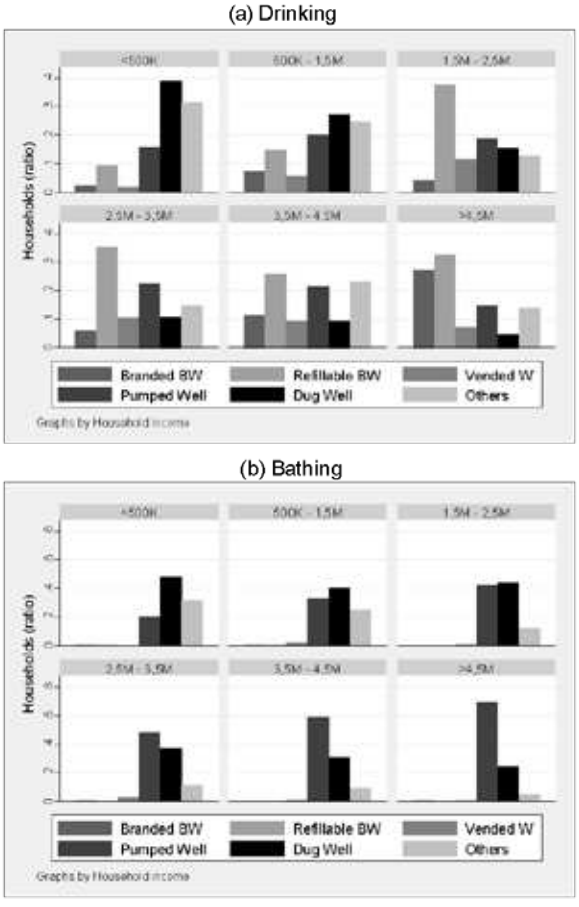


Figure 5: Current behaviour in accessing clean water for (a) drinking and (b) bathing from Water INDII Survey 2017

Table 1: Distance from Main Road

Village	Number of Household	Non-piped Households	Basic water needs (m^3/months)
Siantan Hulu	6,909	1,919	19,190
Siantan Tengah	8,763	5,389	53,890
Siantan Hilir	8,191	5,376	53,760
Batulayang	4,395	3,238	32,380
Total	**28,258**	**15,922**	**159,220**

3.3 How can clean water access for low income households in Pontianak Utara be improved?

Potential water sources that can be utilised as raw water then as clean drinking water include springs, surface water and groundwater. Specific in Pontianak City, there are two sources of water with the spreading that is determined by the potential of each type of source.

First, the potential for surface water under Law No. 7 of 2004 on Water Resources, is all water found on the soil surface. Surface water is water that comes from sources of rivers, lakes, reservoirs, situations, swamps and so on. The potential of surface water in Pontianak City comes from Kapuas River, Landak River and Penepat River [8]. The rivers are very big role for the aspects of community life around them, other than as a source of raw water, irrigation sources, and water transportation.

Other potentials source of water is the groundwater that many residents use as drinking water. Ground water demand from year to year shows an increase that causes an increasing demand for groundwater discharge. Increased demand for ground water with easy procedures without regard to the availability of groundwater present in a region [7],[8].

However, PDAM Tirta Khatulistiwa does not yet get involved intensively for the utilisation of those alternative water sources to obtain the efficient level of production. Clearly there has been some issues to elaborate, such as the assessment, development direction, regulation and protection for sustainability purpose.

Meanwhile, the Integrated Database for the Social Protection Program in 2015 shows that North Pontianak is one of the districts concerned with the low income (Fig.7) and low interest of residents to use water services from PDAM, which is approximately below 40 percent (Fig.6). According to PDAM Tirta Khatulistiwa Pontianak [8], only 55 percent of residents of North Pontianak Sub-district are utilising a tap water connection (Fig.6). North Pontianak sub-district became a sub-district with the smallest number of their customers.

There are several reasons why many people have not used PDAM, namely the lack of network-related socialisation, varying economic levels, to the high utilisation of river water, groundwater and rainwater that are considered cheap[2],[6],[10],[13].

The estimated demand for water in Pontianak Utara district based on equation 1 is 159,220 m3/ month. This number is derived from the basic water needs of 10 m3/ household/ month and number of households without access to piped water (Table 1).

Figure 6: Apart from Siantan Hulu, three other villages have pipeline water access rate below 40 percents

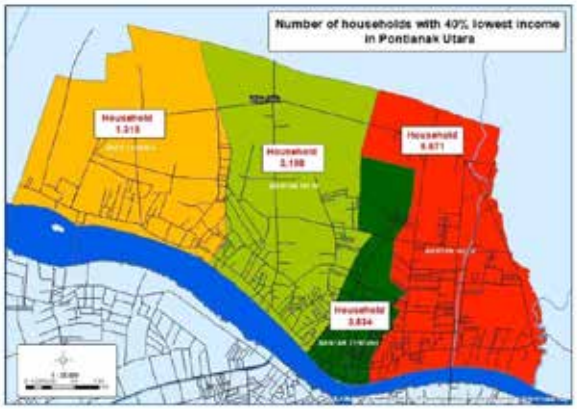


Figure 7: All villages still use river water for laundry and bathing

3.3.1 Water Needs (Wn) for households in Pontianak Utara. The population data of North Pontianak sub-district is sourced from Central Bureau of Statistics [11] which is 124,645 person, and standard of clean water requirement follow Indonesian National Standard (SNI) 03-7065-2005 that is 120liter/person/day. So if substituted in the above formula becomes:

$$Wn = 124.645 persons x 120 liter/day$$
$$Wn = 14,957,400 (liter/day)$$
$$Wn = 14,957 (m/day)$$

Based on the above calculation it is known that the need for clean water at the North Pontianak sub-District (based on the total population of the area) is 14,957,400 (liter/day) or 14,957 (m^3/day). The value is the basis for estimating the amount of Rainwater Harvesting (RH) that must be available to be managed to meet the water needs of the area.

3.3.2 Water Bank with Rainwater Harvesting System (RH) for Alternative Solution. Water storerooms are offices that are empowered to hold water rain for re-utilise (re-utilise). The Rainwater Harvesting System - is activity or endeavor to gather water rain falling on the reservoir above the surface of the earth, either a top of a building, street, yard, and for extensive scale shape water catchment regions. By looking at the condition of the region and the problems that exist in North Pontianak, then the most appropriate method for alternative solutions to apply is Rainwater Harvesting System by utilising the field catchment (Fc) in a top of building to fill Water Bank.

In order to fulfill the water requirement in North Pontianak sub-District area about 14.957.400 (liter/day) or 14,957 (m3/day), the average daily rainfall of the area must be conditioned with the rainfall field catchment(Fc) area, so that it can meet the expected water needs:

$$[RH]14,957(m^3) = [Fc(m^2)]x[Ra]0.03(m/day) \quad (3)$$

$$[Fc(m^2)] = [RH]14,957(m^3)/[Ra]0.03(m/day) \quad (4)$$

$$[Fc(m^2)] = 498,566.7(m^2) \quad (5)$$

The average Rainfall (Ra) value is derived from daily rainfall data processing (mm/day) from online Climate Data Library over the past 14 years (Fig. 8). This data is used to calculate the catchment field relation to the large water demand in North Pontianak. The result of rainfall average by online processing is 30 mm/day or 0.03 m/day.

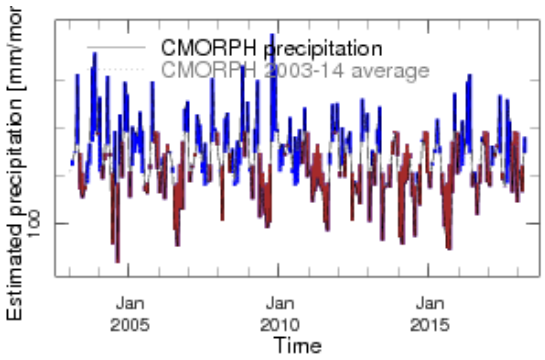


Figure 8: Average of Rainfall (precipitation) in Pontianak 2003-2017

To meet the water requirement of 14,957 m^3, using rainfall average 0.03 m/day, then it needs wide of catchment area about 498,566.7 m2 on Water Bank installation. Since the catchment field for the Rainwater Harvesting System required in the Water Bank installation is too large, there can be several installations in several locations around the community settlement.

If it is assumed that every catchment field used for Rainwater Harvesting is as same wide as the Runoff Harvesting Field catchment area like Embung, that serves as a Water Bank, with a minimum volume of 500 m^3, then, Water needs of North Pontianak divided by Water Bank (embung) capacity volume, so it will require the installation of about 30 Water Bank unit, like a calculation below:

$$14,957m^3/500m^3 = 29.9 \quad unit$$

or equal to 30 unit Water Bank. To fill each unit of Water Bank with a volume capacity of 500 m3, the required Field of catchment (Fc) is:

$$500(m^3) = 0.03(m/day)xFc(m^2)$$

$$Fc = 500/0.03$$

$$Fc = 16.666,7m^2$$

$$Fc = 17m^2$$

So if 30 units of Water Bank installation contains about 500 m^3, then the area of North Pontianak will get water supply from Rainwater Harvesting as big as 15,000 m3, or more than is needed by the local community, i.e. 14,957 m^3.

4 CONCLUSION

Based on the results of research and analysis that have been described above, it can be drawn conclusion as follows:

- The uneven distribution of access to clean water in Pontianak is caused by income inequality;
- Current behaviour in accessing clean water some poor households relies on dug well and others water sources, while richer households choose bottled water and pump well;
- PDAM can build waterbank as a solution for people in the north of Pontianak City to get clean water. Waterbank can be built by analyzing the potential for rainwater harvesting.

REFERENCES

[1] M Anshari, G. Z.and Afifudin, M. Nuriman, and et.al. 2010. Drainage and land use impacts on changes in selected peat properties and peat degradation in West Kalimantan Province, Indonesia. *Biogeosciences, 7, 3403-3419* (2010).

[2] J Anthony Catanese and C.James Snyder. 1994. Perencanaan Kota (translation). (1994).

[3] K. Vermeulen B. Hoitink T. EHuisman, A and M Pramulya. 2016. Observing River Migration Development from LANDSAT Images: Application to The Kapuas River, West Borneo. *Conference Paper* (2016).

[4] E. Gusmayanti, Z. Anshari, G, M. anSholahuddin Pramulya, and N. Kusrini. [n. d.]. Sago palm as a potential crop for peat restoration in West Kalimantan. *International Conference on Biodiversity, Abs Soc Indon Biodiv* 3, 7 ([n. d.]).

[5] E. Gusmayanti and Z.and Pramulya M.and Ruliyansyah A Anshari, G. [n. d.]. Empiric equations to estimate carbon emission from water table level in palm oil plantation cultivated on peatland. 4, 7 ([n. d.]).

[6] A. S. Lestari, Aditiajaya, E. Widianingsih, and H Dharmawan. 2009. Monitoring Kualitas Air Oleh Masyarakat. (2009).

[7] Badan Peningkatan Penyelenggaraan Sistem Penyediaan Air Minum. [n. d.]. Indonesia water supply infrastructure PPP investment opportunities 2017. ([n. d.]). http://kpsrb.bappenas.go.id/data/filedownloadbahan/Bappenas_20170917_Presentation%20Material%20for%20IIIF%20Korea_rev8.pdf

[8] PDAM Pontianak. 2015. Rencana Induk SPAM Kota Pontianak 2015-2035. (2015). http://jdih.pontianakkota.go.id/

[9] M. Pramulya and B Gandasasmita, K.and Tjahjono. 2011. Kajian Geomorfologi Dan Resiko Banjir, Serta Aplikasinya Untuk Evaluasi Tata Ruang Kota Sintang. *Jurnal Ilmu Tanah dan Lingkungan* 13, 2 (2011).

[10] F.and Hoff R V D. Rukmana D.W, N.and Steinberg. 1993. Manajemen Pembangunan Prasarana Perkotaan. (1993).

[11] Badan Pusat Statistik. 2016. Kota Pontianak Dalam Angka Tahun 2016. (2016). https://pontianakkota.bps.go.id/publication/2018/01/05/099d5d0a56df443f82ff3634/kota-pontianak-dalam-angka-2016.html

[12] A.and Hoitink A.and Pramulya M Vermeulen, B.and Huisman. 2016. Migration of banks along the Kapuas River, West Kalimantan. *International Conference on Fluvial Hydraulics, RIVER FLOW* (2016).

[13] Y. Yuliani and M Rahdriawan. 2014. Kinerja Pelayanan Air Bersih Berbasis Masyarakat di Tugu rejo Kota Semarang. *Jurnal Pembangunan Wilayah dan Kota* (2014).

Inferring Energy Consumption towards Urban Development by Combining Social Media Activity Density and Socio-Economics Statistics

Dharma Aryani
Politeknik Negeri Ujung Pandang
Makassar, Sulawesi Selatan
dharma.aryani@poliupg.ac.id

Wini Widiastuti
Badan Pusat Statistik
Mataram, Nusa Tenggara Barat
winiwidiastuti@bps.go.id

Dwi Martiana
Universitas Jember
Jember, Jawa Timur
dmartiana@unej.ac.id

Pamungkas Jutta Prahara
PulseLab Jakarta
Jakarta Pusat, Jakarta
pamungkas.prahara@un.or.id

Muhammad Rizal Khaefi
PulseLab Jakarta
Jakarta Pusat, Jakarta
muhammad.khaefi@un.or.id

ABSTRACT

This paper examines the energy consumption in correlation with social media activity density and socio economic statistics. By analyzing the aggregated Twitter data, the users can be classified into two categories, *local* and *tourist* user. Spatial mapping of district-based data has shown a significant correlation between Twitter activities, energy consumption and socio economics. Additionally, there was a considerable statistical relationship between Twitter activities and energy consumption as the results of Panel Data regression. A further analysis of social media activities and socio-economic statistics by using step wise regression method leads to a far-reaching statistical model to infer the energy consumption in Bali with an accuracy rate higher than 95%. The number of tourists visiting as well as the population density of each district are highly contribute to the changing pattern of electricity consumption in the province of Bali.

KEYWORDS

big data, urban development, social media, socio economics, energy consumption

1 INTRODUCTION

Sustainable urban development has risen high on current global priority since urban areas are multi-functional complex systems. The biggest issue in development priorities is to ensure the reliable infrastructure and services can be provided to support human activities, such as energy supplies, roads, health systems, and accessible clean water. A proper characterization of urban area dynamics and structures would facilitate a better future planning of urban and regional development.

Big data have been used to examined several elements of human behaviour such as localization and mobility of people in different circumstances. In big data frame, information from social media is a favorable reference that could provide location and spatial data in varied range of coverage. Social media data has been used as a proxy to represent the characteristic of human activity in different circumstances. This is very feasible to enrich existing data collection methods for mapping activity patterns and location-based experiences of people. The data could potentially provide continuous information about the users activities and the pattern of their interactions with the environment. Several studies have shown the use of social data to asses people activities and movements in urban environments [12], [1],[10]. The commuting pattern between different places and activities conducting from these areas, influencing human mobility and directly affecting the infrastructure planning in an urban area.

The case study area for the research is the Province of Bali which is administratively divided into 9 regencies (Denpasar, Badung, Gianyar, Bangli, Karangasem, Klungkung, Jembrana, dan Tabanan). The urban development is something that poses every day issues in Bali, from the infrastructure planning, study of the existing urban area properties and forecasting the future needs of urban areas to prepare a better planning. Bali is one of the most popular tourism destinations in the world. Therefore, the social media data should be a very descriptive information to explain about the urban living characteristics.

This research aims to investigate the correlation of tourist tweet activity with the energy consumption pattern and to define a model which represents the relationship between energy consumption with socio-economic statistics. Various correlation analysis are investigated to gain insightful information about the influencing factors of energy consumption. Eventually, it can be validated that the large number of tourists coming to Bali Island as well as the population density of each regency in the province of Bali contributed to the pattern of electricity consumption.

The presentation in this paper is divided into five sections. Section 2 covers the literature review of previous studies. Section 3 presents the dataset and methodology of te research. Results and discussions are elaborated in Section 4, followed by a brief conclusion in Section 5.

2 RELATED WORKS

There has been a number of recent studies to analyze the relationship between social activities as indicators to learn about human mobility behavior by using human mobility data from social media. Several studies use social media check-in data from Foursquare to analyze collective human mobility and activity patterns to infer urban [3, 4].

A significant body of literature from social science, environmental study, computer science, and machine learning investigates new ways to interpret related informations between online and offline interactions [5, 8], urban dynamics in large scale and people behavior in regards to the location technologies effect [9].

Online social networks data which provide information of geography and location in have drawn significant attention. An exploration of temporal dynamics in correlation with on-line social activities has been conducted in [7]. In addition, a model to predict the location of users has also been identified using a spatial distribution of words in Twitter user-generated content [2]. Modeling of network properties has been examined in association with local geography [11]. Extended study about the pattern of location sharing services used by users, and its privacy issues has been studied in [9].

The social media dataset used in this research is collected from Twitter. Studies have proven that Twitter data and specific contextual information might serve as an indicator on how strongly the virtual and physical worlds are connected with each other. In online social networks like Twitter, users create an online profile and communicate with other users by sharing common ideas, activities, events or interests [6]. Each tweet contains a corresponding geography location that is collected from the GPS sensor within the mobile device. Thus, any post by users contains a spatiotemporal data (geolocation and timestamp of tweet) and a semantic information from the tweet message [10]

3 RESEARCH METHODOLOGY

3.1 Data Sources

In accordance with the research task to infer the energy consumption based on social media activity and socio-economics, this research utilizes three groups of dataset. A data log of Twitter activities in the Province of Bali is used as a proxy for further scientific analysis.The Twitter data is in 30 minutes data aggregation for the year of 2014. For electricity information, the best reference is the one which is provided by PT. PLN as the institution that manage energy operation, marketing and distribution in Indonesia. PLN data consists of the peak load of electricity demand in each of power distribution substation in the Province of Bali. This PLN data is in two period aggregation, only day and night peak load variable. The location map of distribution substations is presented in Figure 1. The other data set is socio-economics statistics from Badan Pusat Statistik (BPS) which provides annual basis data of population, regency area, and poverty status.

3.2 Statistical Analysis Method

The pre-treatment procedures are taken to accommodate the data limitations. Firstly, since the Twitter data are in 30 minutes aggregation while the PLN data are in Day and Night aggregation, they are incomparable. Therefore, the initial process is re-aggregation of Twitter data into two time interval (Day and he Night). Secondly, excluding all datasets in January 2014 in regards to the unreliable situations on Twitter data generation during the period. So that, only the eleven (11) months of data log that can be used in the research.

Figure 1: Distribution substation

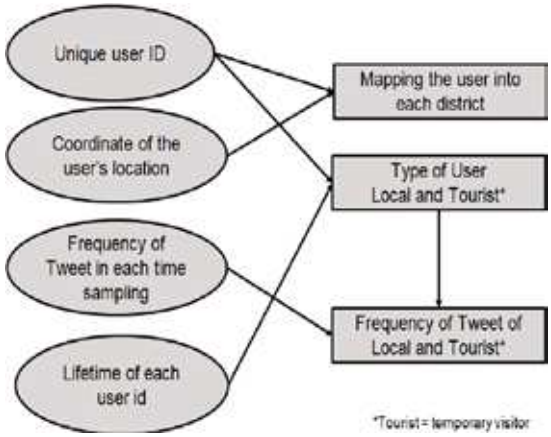


Figure 2: Data selection of Twitter dataset

In addition, the peak load data of electricity consumption are converted into the power consumption variable in KiloWatt per Hour(KWh) which is commonly used as a variable to describe the amount of electricity consumption. Figure 3 shows four entries of Twitter data log that are used in this study. They are *unique user ID, lifetime period, location coordinate, and frequency of tweet.* Initially, the *unique user ID and location coordinate* data are taken for mapping the users into each regency. Furthermore, based on the *lifetime period of each user*, the type of users are classified into two categories, *local* or *tourist*. The last parameter from the Twitter data is frequency of tweet from *local* and *tourist* users.

The classification of twitter user types is determined from *lifetime period* data of each user. This procedure is realized by analyzing the histogram plot of Twitter activities in daily *lifetime period* in regency area. The histogram data is presented in Figure 3. In order to differentiate the *local* and *tourist* users, it is assumed that tourists are temporary visitors including from domestic or international who will only stay in a short term period. Therefore, a *quantile* analysis is carried out to have a representative cutting point for data classification. The result of quantile analysis is that the 75%

Figure 3: Histogram of Twitter users based on daily lifetime period in all regencies

value is on the 7^{th} days of the histogram plot. Thus, it is concluded that all users who regularly using Twitter in seven days can be categorized as *tourist*, and the users who stay active after the 8^{th} onwards are defined as *local* users.

The next statistical processing is calculation of population density of local people and the % of poverty in each regency. These variables are utilized as the model parameter in illustrating the correlation between socio-economics condition and the energy consumption in urban areas. The correlation analyses are implemented in several approaches. Started from basic linear correlation, panel data regression, and step-wise regression analysis. The results of these statistics evaluation are elaborated in Section 4.

4 RESULTS AND DISCUSSIONS

4.1 Correlation between social media activities and energy consumption

A set of proxy data for the correlation analysis is arranged by employing data from Section 3. where the Twitter users are differently identified as *local* or *tourist* users based on the lifetime period of user account. An informative spatial mapping is presented in Figure 4 to provide detail information of the classification results of Twitter users by regency in the Province of Bali. In general, at the province level, the number of total *tourist* users are almost 50% higher than the total local users. Another interesting finding that draws attention from the figure is the location preference for the tourist visit or stay. It can be seen that in 2014, the tourist visit centralized in four regencies, i.e. Badung, Denpasar, Gianyar, and Tabanan. Surprisingly, these regencies are indeed the tourism destinations in Bali where most of the prominent tourism sites, attractions and are located there. This finding also supported by the facts that accommodation for tourist or temporary visitors are centralized in these four regencies. Therefore, it is clear that the proposed approach to classify the tourist users based on *lifetime period* from Twitter dataset is a reliable approach to get a picture of tourism localization pattern in the province.

Furthermore, A correlation analysis is carried out to find the relationship model between energy consumption and social media activities. The results are summarized in Table 1. It can be inferred that the number of twitter users and the frequency of tweeting activities in each regency are positively correlated with the energy consumption.

Table 1. Correlation analysis between social media activities and electricity consumption

Correlation Coefficient	Electricity Consumption
Number of tourist	0.94
Number of local	0.89
Tweet freq.of tourist	0.83
Tweet freq.of local	0.99

Panel Data Regression

At this stage, correlation between the number of Twitter users and electricity consumption in the province of Bali is examined using a panel data regression. This approach is deployed to accommodate the regression analysis for functions which has combined input datasets, in cross section and time series. The cross section data in this research the number of twitter users and electricity consumption per regency in the Province of Bali, consist of 9 regencies. The time series data is the number of twitter users and electricity consumption per month in the period of February to December for the year of 2014. In this research, the identified model is a function $y = f(x)$, by defining y as the electricity consumption, and x as the number of Twitter users.

A basic panel data regression model is given as

$$y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it} \tag{1}$$

where y is the dependent variable, x is the independent variable, α and β are model coefficients, i and t are the indices for individuals and time, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$. The ϵ_{it} is independent, identical in normal distribution, mean 0, variance σ^2

The intercept α_i is estimated in Fixed Effect Model approach. In this case, each data is analyzed based on the time difference and the regency difference, using the dummy variable technique that is accommodated by the difference of intercept between observations. The estimation result of the regression model above is,

$$y_{it} = 18201 + 20.3x_{it} \tag{2}$$

The regression model correlating the number of twitter users with electricity consumption is significant at $\alpha = 0,05 (p < 0.0001)$ with the coefficient of determination (R^2) for 92.4%. Correspondingly, Fixed Effect Model analysis is given by considering the time series data (February to December, $T_1, T_2, \ldots . T_{11}$) expressed in the form of dummy variable. However, since T_{11} is highly correlated with variable x, this data is excluded from the equation. Regression function is written as follows,

$$\begin{aligned} y_{it} = {} & 54184 + 20.5x_{it} - 81881T_1 - 40636T_2 - 43691T_3 - 73496T_4 \\ & - 60340T_5 - 58326T_6 - 71380T_7 - 25908T_8 + 4312T_9 + 32307T_{10} \end{aligned} \tag{3}$$

The regression model in Eq. 3 has a $\alpha = 0.05 (p < 0.0001)$ with a slightly higher coefficient of determination (R^2) for 92.7%.

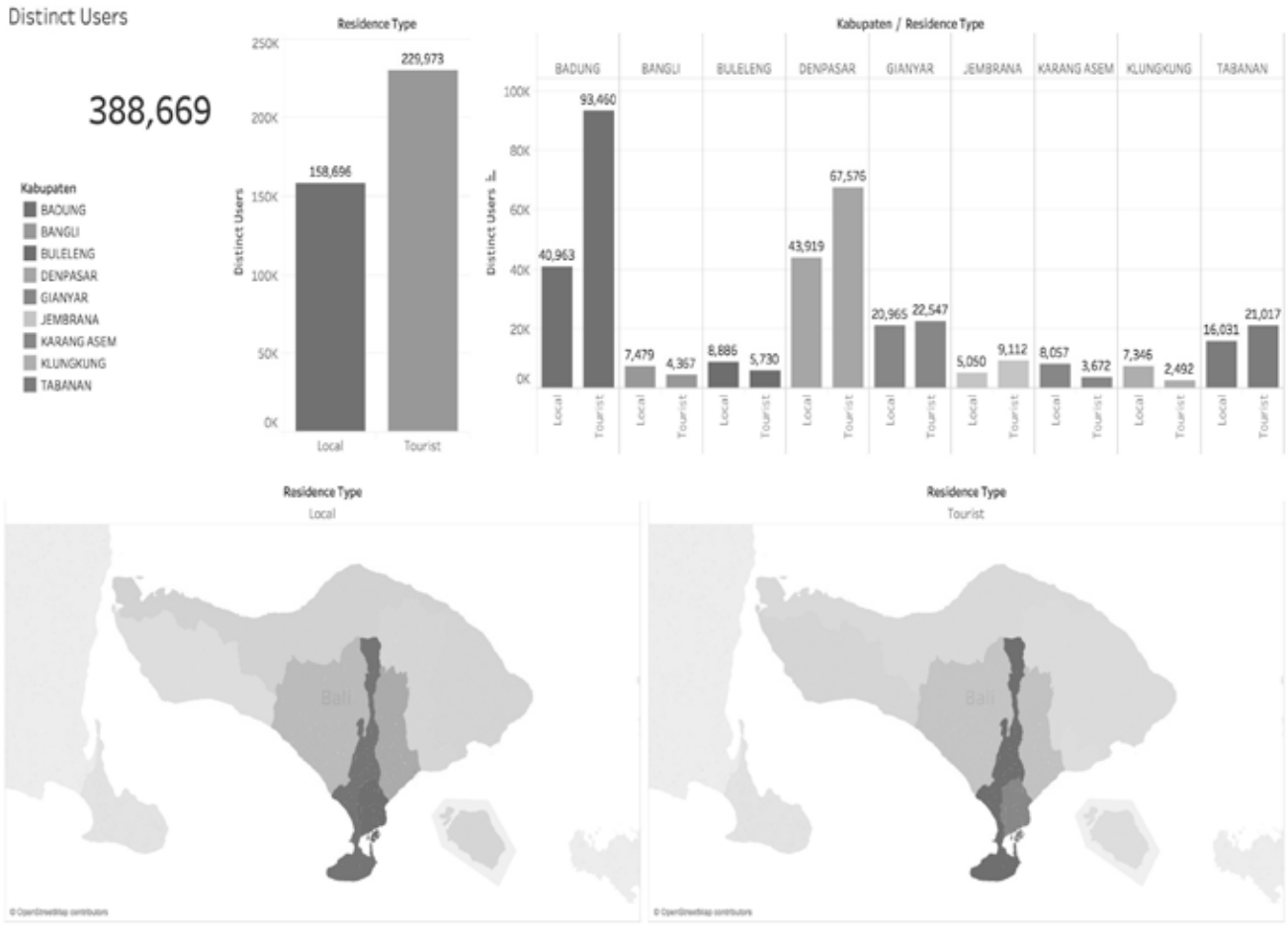


Figure 4: Local and tourist-defined users based on Twitter *lifetime* data

The last FE Model analysis is carried out by considering the regency data ($D_1, D_2, \ldots..D_9$) are the dummy variables. D_9 is also excluded from the regression equation. The R^2 value increased significantly to 99.5% from the estimated model which is formulated in the regression function below„

$$y_{it} = 163785 - 4.92x_{it} + 1092701D_1 + 750835D_2 + 57323D_3 - 51704D_4 - 147860D_5 - 4830D_6 - 51700D_7 - 86750D_8 \quad (4)$$

4.2 Correlation between social media activities, socio economics and energy consumption

It is believed that a spatial mapping is a descriptive way to examine the energy consumption in urban area as corresponded to social media activities and the socio economic condition. The social media activities is represented by the number of overall Twitter users , while the proxy of socio economic statistic is population density. A regency-based mapping is visualized in Figure 5 to infer the energy consumption towards urban development. It can be analyzed that as the population density increases in an urban area, the energy consumption shows the same pattern to level up. The same relationship also appears from the perspective of social media dynamic, the denser the area, the higher energy level is needed. Interestingly, there is a situation where the population density and twitter users are low but its energy demand is higher. A simple analyses is that regency is an industrial area with most of the consumers are in industrial scales.

Inferring the energy consumption using the socio economics data is proceeded by taking the value of population density and the level of poverty in each regency. The comparison of annual pattern of electricity consumption and population density is captured in Figure 6.

The graph shows that the overall trend of electricity consumption and population density through regencies are almost similar, when the population density is high, the energy consumption is also high. However, in Badung, the population density is just above Gianyar, but the electricity consumption is far higher. This reality is driven by condition where Badung is the center of tourism activities and

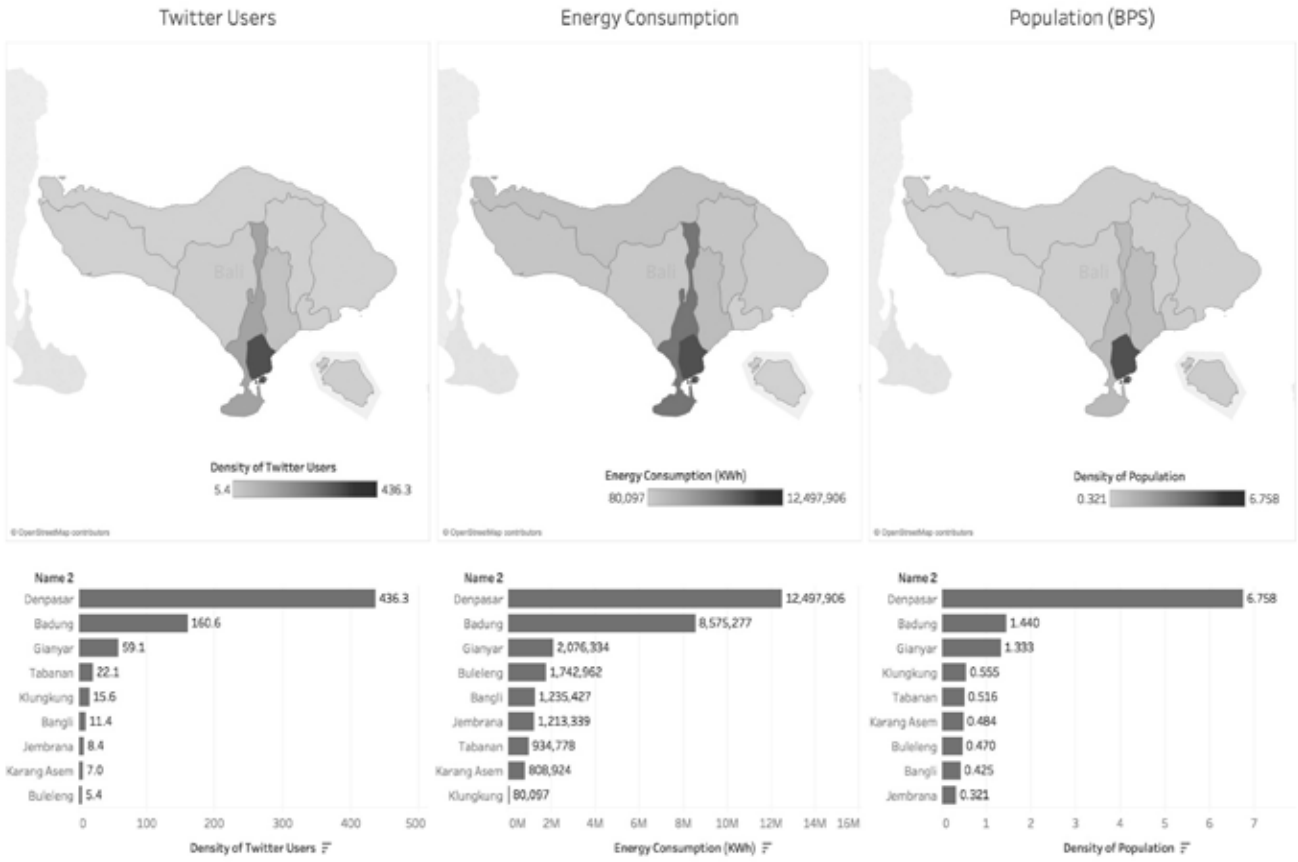


Figure 5: Spatial mapping of Twitter users, energy consumption, and population density

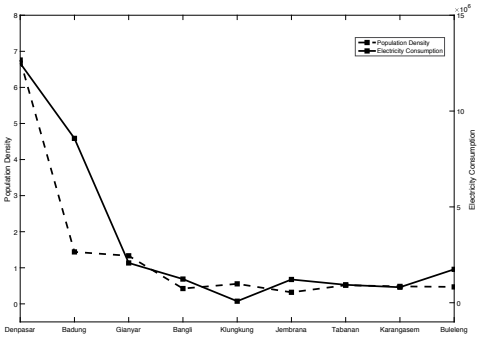


Figure 6: Regency-based electricity consumption and Population Density

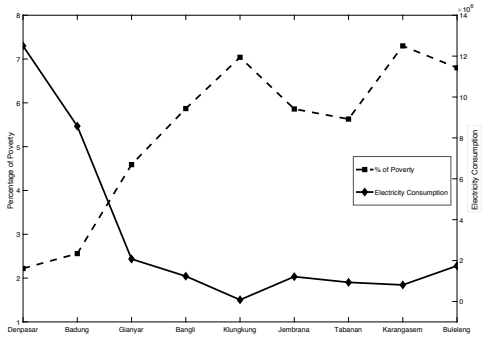


Figure 7: Regency-based electricity consumption and % of Poverty

accommodations which means that the energy consumption is highly correlated with the number of tourists.

Furthermore, relationship between the % of poverty and the energy consumption by regencies can be evaluated from Figure 7. It is very clear that these two variables are significantly correlate to each other in an inverse correlation. The electricity consumption decreased when the poverty level increased, this mush illustrates that the economic condition is the basic factor which determines

the level of energy consumerism. Table 2 provides the correlation coefficient between socio economics statistics and the energy consumption.

Table 2. Correlation analysis between socio economics and electricity consumption

Correlation Coefficient	Population Density	% of Poverty
Electricity Consumption	0.88	0.91

Step-wise Regression

For the purpose of socio economic data analysis, step-wise approach is selected as the regression method. Stepwise regression analysis is a combination of forward and backward methods. It is employed to obtain the best model to illustrate relationship between several input variables, as well as get the right predictor for an output. In this step, the input variables are the number of local and tourist users of Twitter , average of daily consumption expenditure, population density and the percentage of poverty in all regencies.

The principle of stepwise regression is including variables that have the highest correlation and significant with the dependent variable, in this case the average electricity consumption per region in the province of Bali. The next variable to be included in the model is the one with highest partial correlation and significant with the dependent variables. After all variables have been formulated in the model, it is evaluated to exclude the insignificant variables from the model.

It is found that the average of daily consumption expenditure and the percentage of poverty are insignificant with p-value ≥ 0.2. Therefore, it is excluded from the model and it leads to a regression function of electricity consumption (y) as in Eq. 5 with only three inputs, the Twitter users (x_1 =*local* and x_2 =*tourist*) and x_3 = population density of each regency. The p value is 0.0001 with coefficient of determination (R^2) reaches 97.43%.

$$y = 792665 - 1719823x_1 + 1338962x_2 + 1502401x_3 \quad (5)$$

However, a partial test of all the model coefficients has shown that the number of *local* users have p value= 0.15, so that this variable is removed from the model estimation on the next step wise regression. In the second regression, only the number of *tourist* users and population density data are taken for the regression. The regression result is formulated as follows,

$$y = -1711817 + 7371491x_2 + 1115017x_3 \quad (6)$$

The p value is 0.00001 with coefficient of determination (R^2) reaches 96.66%. These findings are evident that the electricity consumption in the province of Bali is determined by the number of twitter users of the tourist group and population density.

5 CONCLUSIONS

This research has proven that social media data can potentially be used to differentiate the type of user as local or tourist. By combining different type of statistical analysis , it can be clearly inferred that energy consumption is significantly correlated with social media activities and socio economic statistic. To sum up, the large number of tourists coming to Bali Island as well as the population density of each regency in the province of Bali contributed to the pattern of electricity consumption.

ACKNOWLEDGMENTS

The authors would like to acknowledge PulseLab Jakarta for the initiation of Research Dive 6 "Urban and Regional Development".

REFERENCES

[1] Gennady Andrienko, Natalia Andrienko, Harald Bosch, Thomas Ertl, Georg Fuchs, Piotr Jankowski, and Dennis Thom. 2013. Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering* 15, 3 (2013), 72–82.

[2] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 759–768.

[3] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. 2011. Exploring millions of footprints in location sharing services. *ICWSM* 2011 (2011), 81–88.

[4] Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. (2012).

[5] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. 2010. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 119–128.

[6] Nicole B Ellison et al. 2007. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication* 13, 1 (2007), 210–230.

[7] Scott A Golder, Dennis M Wilkinson, and Bernardo A Huberman. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Communities and technologies 2007*. Springer, 41–66.

[8] Eric Gordon and Adriana de Souza e Silva. 2011. *Net locality: Why location matters in a networked world*. John Wiley & Sons.

[9] Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. 2011. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2409–2418.

[10] Enrico Steiger, René Westerholt, Bernd Resch, and Alexander Zipf. 2015. Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems* 54 (2015), 255–265.

[11] Sarita Yardi and Danah Boyd. 2010. Tweeting from the Town Square: Measuring Geographic Local Networks.. In *ICWSM*. 194–201.

[12] Weiyang Zhang, Ben Derudder, Jianghao Wang, Wei Shen, and Frank Witlox. 2016. Using location-based social media to chart the patterns of people moving between cities: The case of Weibo-users in the Yangtze River Delta. *Journal of Urban Technology* 23, 3 (2016), 91–111.

http://rd.pulselabjakarta.id/

Australian Government

PULSE LAB JAKARTA

Pulse Lab Jakarta is grateful for the generous support from
the Government of Australia