Seminar Component

Name of authors:

Titi Kanti Lestari; Siim Esko; Alexander Rayner, Amalia Adininggar Widyasanti

Organizations

BPS-Statistics Indonesia

Positium LBS

Amadeus IT Group S.A.

Ministry of National Development Planning/Bappenas, Indonesia

Contact address

BPS-Statistics Indonesia

Jl. Dr. Sutomo No. 6-8

Jakarta 10710 Indonesia

Contact Emails :
titi@bps.go.id; siim.esko@positium.com; alex@smartdata.travel; winny@bappenas.go.id

Title of Paper: The use of Big Data as the Leading Indicator of Tourism Demand

Abstract

Tourism has become a national priority program in Indonesia's mid-term National Development Plan. Effective policy making requires the most accurate and real time data when possible. BPS-Statistics Indonesia publishes the official tourism statistics for Indonesia, using data sources that are mainly administrative (Immigration data) that have limitations. In the Big Data era, other data about tourism destinations from big data sources such as Amadeus Destination Insight (Amadeus Data), and Mobile Positioning Data (MPD) can be used in conjunction with administrative data and as leading indicators of tourism demand. This paper aims to analyse and compare the patterns of tourism demand in Indonesia from Amadeus data and MPD with tourism statistics from BPS-Statistics Indonesia. This paper will evaluate the prediction results by comparing several time series models, and furthermore, it will compare and correlate the Amadeus data and MPD with official data. It will explore whether Amadeus data and MPD have correlation with official data, if accurate predictions can be produced, and if they can be used as leading indicators of tourism demand for Indonesia. This paper aims to help improve the quality of data for decision-making in the tourism sector, especially with infrastructure and development, to facilitate tourism demand in Indonesia and ultimately optimise the contribution of tourism to Indonesia and the SDGs.

Keywords: tourism demand, Amadeus, Amadeus Destination Insight, leading indicators, Bayesian, big data, Mobile Positioning Data, MPD

Table of Contents

1. Introduction

As the 17 Sustainable Development Goals (SDGs) and the corresponding 169 SDG targets offer the world a new direction, tourism can and must play a significant role in delivering sustainable solutions for people, the planet, prosperity and peace.

Tourism has the potential to be an economic powerhouse, as it is the third highest world category in export earnings in 2015, representing 10% of world GDP, 30% of services exports and contributing to 1 out of every 10 jobs in the world. According to the United Nations World Tourism Organisation (UNWTO), International Tourist Arrivals increased by 6% to 1.4 billion in 2018, and are expected to reach 1.8 billion by 2030, which means on average 5 million people will be crossing international borders every day, and with the growth of the global middle class these numbers are likely to be exceeded.

Tourism has potential to contribute, directly or indirectly to all of the SDGs goals. In particular, it has explicitly been mentioned as targets in **Goals 8, 12 and 14** on inclusive and sustainable use of oceans and marines resources, respectively. By 2030, Target 8.9 sets out to devise and implement policies to promote sustainable tourism that creates jobs and promotes local culture and products. Furthermore, by 2030, Target 14.7 aspires to increase the economic benefits to small island developing States and least developed countries from the sustainable use of marine resources, including through sustainable management of fisheries, aquaculture and tourism. Target 12.b aims to develop and implement tools to monitor sustainable development impacts for sustainable tourism, which creates jobs, promotes local culture and products. The development of new data sources and tools is indeed expected to boldly contribute to the SDGs target 12.b; as a means of SDGs implementation.

Many governments and destinations have recognised the important economic and social impact of International tourists, consequently competition between destinations to attract tourists is extremely strong and continues to intensify.

Every stage of the travel journey from dreaming, planning, booking, experience, to sharing, creates an abundance of data and a digital footprint of visitors. Consequently, an emerging strategy by National Tourism Organisations (NTOs) is to use data metrics and analytics for decision making, enabling marketing to be more targeted and focused to attract high value adding visitors, with a shift away from using the number of visitor arrivals as a performance indicator, to more meaningful indicators such as visitor expenditure, length of stay, and number of jobs in tourism.

As the Big Data era becomes more complex, and the velocity of data continues to accelerate, the role of technology will become critical to making sense of data, understanding data, to make data-driven decisions, and to make them quickly to gain competitive advantage or merely to maintain existing customers.

Attracting target customers from over 1.4 billion annual International tourist arrivals, who are spending over US$1.4 trillion, requires focused marketing and decision making based on data metrics and analytics.

Tourism in Indonesia

Tourism is very important to Indonesia, as it is one of the main economic drivers to the Indonesian economy. Furthermore, tourism is also contributing to the inclusive growth and has potential to uplift the roles of Small and Medium-sized Enterprises to economic activities, as the SMEs can link to tourism business activities.

According to WTTC report (2019), travel and tourism contribute as much as 6.0% to Indonesian GDP, and its growth in 2018 reached 7.8%, far above the Indonesian economic growth in 2018, which was only 5.1%. Furthermore, travel and tourism in Indonesia also contributed to employment as much as 10.3%, and represents approximately 12,900 jobs in 2018.

WTTC Report (2019) also mentions that travel and tourism has significantly contributed to women empowerment and women participation in the economy, as its research found that almost twice as many female employers were in travel and tourism than other sectors. The sector provides more opportunities for women to participate in the workforce, leadership development, entrepreneurship and empowerment than many other sectors, particularly in developing countries.

Tourism development has been set as a priority by the Government of Indonesia in these recent years. The government is planning to spread tourism footprint across Indonesia by developing 10 new Bali destinations, due to their tourism potentials and the aim to increase the role of tourism to the Indonesian economy. The 10 new Bali are Toba Lake, Tanjung Kelayang, Tanjung Lesung, Pulau Seribu, Candi Borobudur, Mandalika, Gunung Bromo, Wakatobi, Labuan Bajo, and Morotai, located across the country. All these priority tourism destinations have been prioritized by the government to be developed in the near future in order to contribute and spread out tourism that leads to the achievement of the Sustainable Development Goals.

Managing tourism development in Indonesia needs to be accompanied by the development of data and tools to monitor the progress and provide a more accurate forecasting tool for policy making process. In this regard, this paper focuses on developing new data sources – big data – to measure the potential demand of tourism as a useful tool for policy makers in Indonesia.

2. Use of big data as the leading indicator of tourism demand

The use of big data as one of data source for official statistics has been recommended by UN Statistical Commission since 2014, as stated in the UN Secretary General Data Revolution Report "A World that Counts". For BPS-Statistics Indonesia, the use of big data is in line with BPS Strategic Planning and Transformation Program. It is believed that statistics based on big data can be compiled automatically, in some cases almost in real time and requires less manual labor. For BPS Statistics Indonesia, the use of big data could reduce respondent and work burden.

The leading indicator approach to economic and business forecasting, pioneered by the National Bureau of Economic Research (NBER), is now widely used in predicting turning points of business cycles in many countries. This method has been used widely due to three reasons. Firstly, early detection and timely recognition of business cycle turning points is important as it would allow policymakers to trigger pre-emptive countercyclical policy measures, businesses to adjust their sales or investment strategy, and investors to reallocate assets among alternative investments to optimize their return. Secondly, it has long been recognized that procedures for making quantitative forecasts, such as standard macroeconometric models, are not appropriate for making turning point predictions that involve detecting regime shifts (Samuelson 1976). Thirdly, since its birth, the leading indicator approach has maintained its standing as a reliable and inexpensive forecasting tool quite successfully. Constructing leading indicators of business cycles requires high frequency data, typically on a monthly basis and a long time series.

Therefore, the availability of big data such as Amadeus data and mobile positioning data prove an opportunity to use it as leading indicator for tourism demand.

3. Data Sources

Immigration Data

BPS mainly uses immigration data for its tourism official statistics on international visitor arrivals. Since 2018 BPS Statistics Indonesia has a memorandum of Understanding with Ministry of Law and Human Rights, which enables the Director General of Immigration to have a technical agreement and provide data to BPS. Immigration data is available on the daily level and detail raw data since 2015 until today.

Amadeus Destination Insight (Amadeus Data)

Amadeus is a leading travel technology company that has over the past 31 years become the world's largest travel data processor, with 8.6 billion travel transactions every day.

Amadeus Destination Insight makes available the most comprehensive air travel data in several modules with data about demand, airfares, destination shopping behaviour, air connectivity, travel agent bookings using the world's 3 Global Distribution Systems (GDS), Amadeus, Sabre, and Travelport. This data is also available for 12 months into the future, enabling forecasting.

This research will use return GDS bookings for the period 1 July to 31 August 2018 and compare with Immigration data and MPD data. By selecting return bookings only this isolates tourists that originate from the same airport that they return to indicating that they reside in this location.

Mobile Positioning Data (MPD)

Mobile Positioning Data (MPD) is provided from mobile network operators and enables insight about the movement and flow of tourists based on their mobile devices.

Mobile positioning data will also be compared with immigration data. BPS uses mobile positioning data from the largest mobile network operator for travel statistics. Mobile positioning data to predict the pattern of international tourism demand was extracted from 5 July 2019 to 31 August 2019. MPD used in the paper was only from one operator and is not extrapolated to reflect the roaming market shares of all operators.

Theoretical Comparison of Data Sources

Definitions – In tourism statistics, country of origin is defined through residence. Country of origin in immigration data is based on passport information. In Amadeus and MPD the country of origin is a proxy of residence. As long as people live in the country of passport, these concepts will be aligned. However, this might not be the case for all people.

Coverage – Immigration data is complete for all official gates of entry to the country. However, immigration data does not have origin information about Indonesians returning on a visit from their residence abroad. Amadeus data includes all return GDS bookings from the country of origin, but excludes bookings direct to airlines. MPD includes all mobile users roaming in the mobile network operator whose data we use, but excludes roamers in other operators.

Timeliness – All the data sources are available within a month after the month end or even before.

Sustainability of data provision – For immigration data, BPS has an MoU with the data provider, making sure the data delivery is sustainable. For Amadeus and MPD, data availability is subject to reaching a contractual agreement with the data providers. There is a threat that the companies stop selling their data, but the larger threat is the decision not to buy the data by the data user. For Statistical Office, once we use data source we have to use it forever for consistency, unless we got another data source that can replace it.

4. Methodology

Bayesian Structural Time Series (BSTS) Model

Bayesian estimation and inference has a number of advantages in statistical modelling and data analysis, for instance, the Bayes method provides confidence intervals on parameters and probability values on hypotheses that are more in line with commonsense interpretations (Congdon, 2006). Andy Pole, Mike West, and Jeff Harrison (1994) stated that Bayesian statistical analysis for a selected model formulation begins by first quantifying the investigator's existing state of knowledge, beliefs, and assumptions. These prior inputs are then combined with the information from observed data quantified probabilistically through the likelihood function – the joint probability of the data under the stated model assumptions. The resulting synthesis of prior and likelihood information is the posterior distribution or information. The posterior is proportional to the prior and the likelihood,

$posterior \propto prior \times likelihood$

the prior to posterior process is referred to as Bayesian learning.

Peter Congdon (2006) stated that the learning process involved in Bayesian inference is one of modifying one's initial probability statements about the parameters before observing the data to updated or posterior knowledge that combines both prior knowledge and the data at hand. Bayesian models are typically concerned with inferences on a parameter set $\theta(\theta_1, \dots, \theta_d)$ of dimension d. Prior knowledge about the parameters is summarized by the density $p(\theta)$, the likelihood is $p(y|\theta)$, and the updated knowledge is contained in the posterior density $p(\theta|y)$. From the Bayes theorem

$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$,

where the denominator on the right side is the marginal likelihood $p(y)$. The latter is an integral over all values of θ of the product $p(y|\theta)p(\theta)$ and can be regarded as a normalizing constant to ensure that $p(\theta|y)$ is a proper density. Then, Bayes theorem can be expressed as

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Brodersen et.al. (2015) stated that structural time series models are useful in practice because of its flexibility and modularity. Structural time series models are state space models for time series data and can be defined as (Brodersen, et.al., 2015)

$y_t = Z_t^T \alpha_t + \varepsilon_t$ as observation equation,

$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t$ as state equation,

where $\varepsilon_t \sim N(0, \sigma_t^2)$ and $\eta_t \sim N(0, Q_t)$ are independent of all other unknowns.

The components of state used in this paper are:

1. Local Linear trend

The local linear trend model is a popular choice for modelling trends because it quickly adapts to local variation, which is desirable when making short term predictions, as such predictions often come with implausibly wide uncertainty intervals (Brodersen, 2015). Furthermore, Kay H. Brodersen (2015) explained that there is generalization of the local linear trend model where the slope exhibits stationarity instead of obeying a random walk, this model can be written as:

$\mu_{t+1} = \mu_t + \delta_t + \eta_{\mu,t}$,

$\delta_{t+1} = D + \rho(\delta_t - D) + \eta_{\delta,t}$,

where the two components of η are independent.

2. Seasonality

The most frequently used model in the time domain is (Brodersen, 2015)

$$\gamma_{t+1} = -\sum_{s=0}^{S-2} \gamma_{t-s} + \eta_{\gamma,t} \,,$$

Where S represents the number of seasons and γ_t denotes their joint contribution to the observed response y_t.

5. Empirical Evidence

Comparison of Data between Amadeus, MPD, and Immigration

In order to investigate whether Amadeus data can show the pattern of international visitor arrival in Indonesia or not, the next step was by comparing the error model resulted from the prediction of data from Amadeus, MPD, and data from Immigration Offices.

Prediction of International Visitors from France to Jakarta

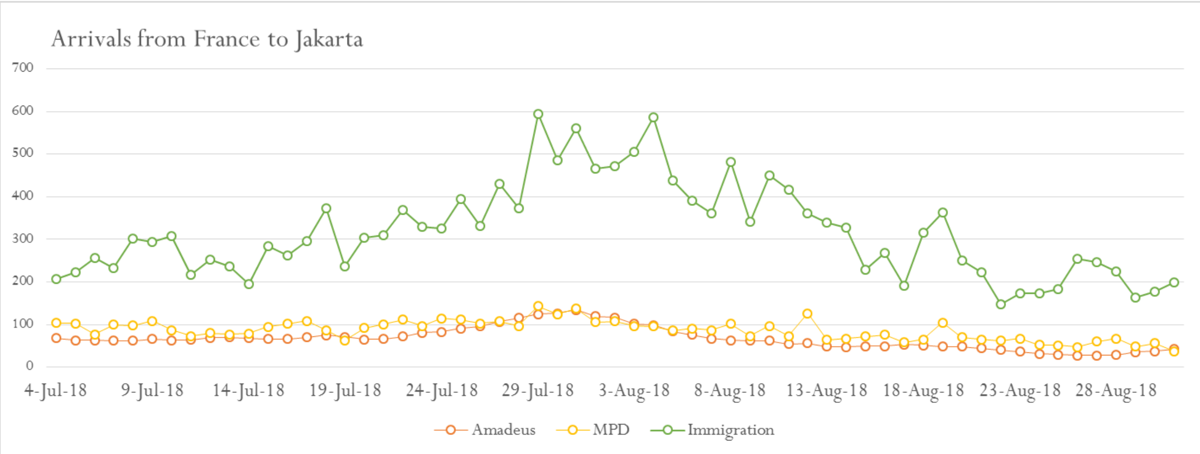

Figure 1. Comparison of arrival data from Immigration, Amadeus and MPD from 5 July to 31 August 2018, France to Jakarta

Figure 2 shows the comparison among Model 1, Model 2, and Model 3 in predicting international visitor arrival in Indonesia. Model 1 resulted from Amadeus data, Model 2 resulted from MPD, while Model 3 resulted from Immigration Offices. Although it is predicted that Amadeus data will increase slower than MPD and Immigration data, Figure 2 also shows that the three models have similar pattern during the period of observation.

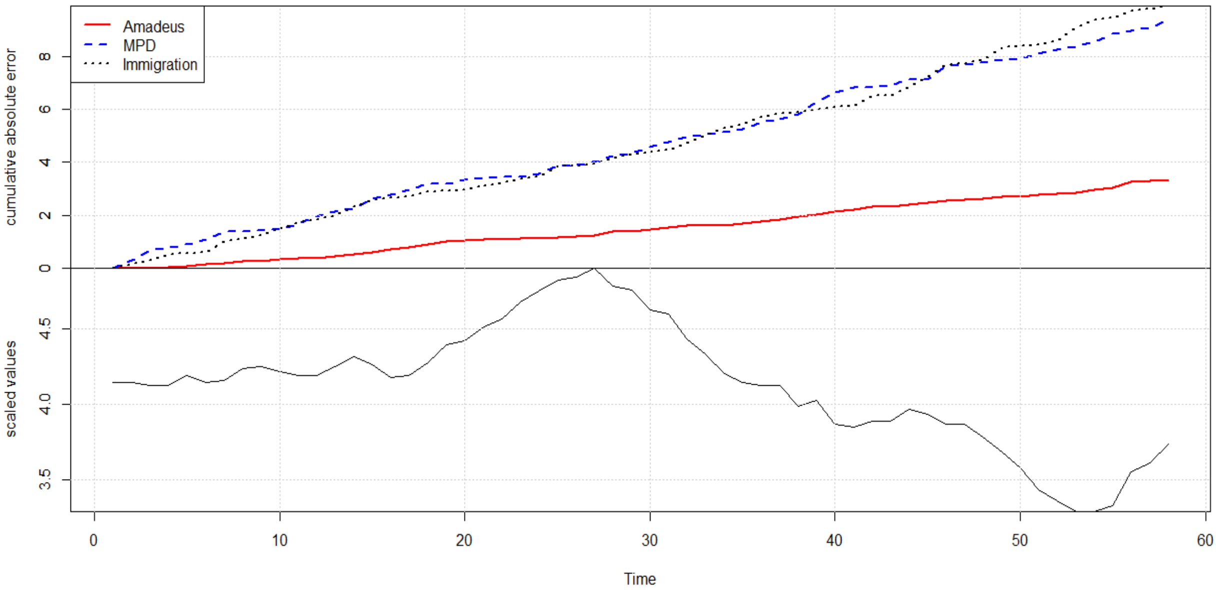

Figure 2. Prediction of International Visitors from France: Comparison of Error from Model 1 (Amadeus Data), Model 2 (Mobile Positioning Data), and Model 3 (Immigration Data)

Table 1. Summary of Model 1, Model 2, and Model 3 for France Visitors

Model	Residual.sd	Prediction.sd	R Square
(1)	(2)	(3)	(4)
Model 1 (Model of Amadeus Data)	0.0320629	0.071944	0.9936448
Model 2 (Model of MPD)	0.1644638	0.200726	0.6736882
Model 3 (Model of Immigration Data)	0.1740144	0.211759	0.7463723

Table 1 is a summary of BSTS model for predicting international visitor arrival in Indonesia. The smallest posterior mean of the residual standard deviation parameter (residual.sd) is model of Amadeus data. In addition, the smallest standard deviation of the one-step-ahead prediction errors (prediction.sd) is also model of Amadeus data. R-square of Amadeus data model is 0.9936448 which means that the fit explains 99.36 percent of the total variation in the data about the average. Meanwhile, R-square of Model 2 (MPD) shows that the fit can explain 67.37 percent of the total variation in the data about the average and Model 3 (Immigration data) can explain 74.64 percent of the total variation.

Table 2. Correlation between Amadeus Data and Immigration Data, MPD and Immigration Data visitors from France to Jakarta

Data	Correlation
(1)	(2)
Amadeus Data and Immigration Data	0.7749745
MPD and Immigration Data	0.7088495

The correlation between Amadeus data and official data is sufficient, as shown in Table 2. International visitor arrival from Amadeus data has 0.77 correlation value with official data. Likewise, the correlation between MPD and official data is also good which accounts for 0.71. It indicates that to some extent Amadeus data and MPD can be used as complement of official data in order to see the pattern of international visitor arrival from France to Jakarta.

Table 3. Correlation between Amadeus Data and MPD

Data	Correlation
(1)	(2)
Amadeus Data and MPD	0.7692177

Prediction of International Visitors from the USA to Jakarta

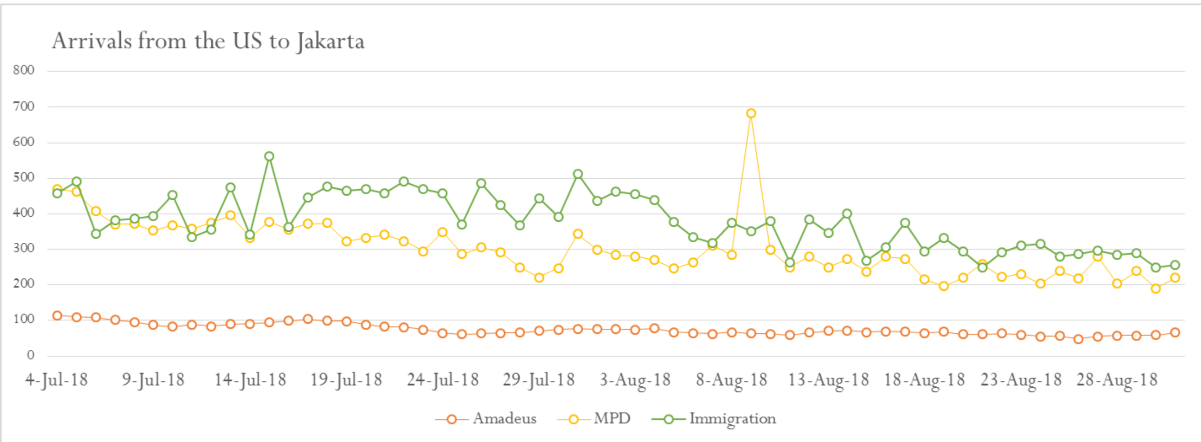

Figure 3. Comparison of arrival data from Immigration, Amadeus and MPD from 5 July to 31 August 2018, US to Jakarta

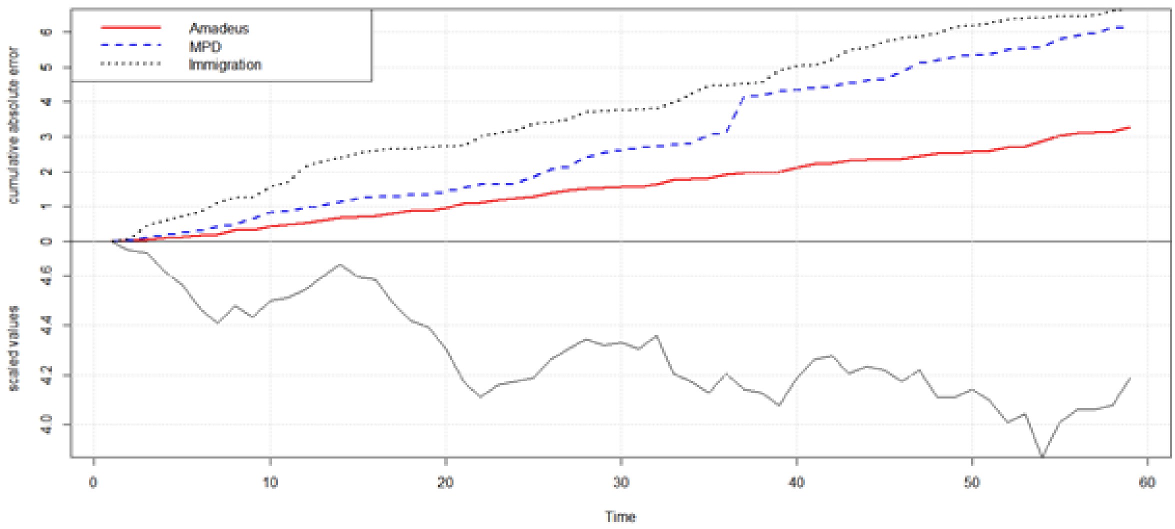

Figure 4. Comparison of Error from Model 1 (Amadeus), Model 2 (MPD), and Model 3 (Immigration)

Table 4. Summary of Model 1, Model 2, and Model 3 for USA Visitors to Jakarta

Model	Residual.sd	Prediction.sd	R Square
(1)	(2)	(3)	(4)
Model 1 (Amadeus Data)	0.0285071	0.0677584	0.9803733
Model 2 (MPD)	0.1618806	0.1698134	0.5642266
Model 3 (Immigration Data)	0.1258762	0.1521643	0.6407415

Table 4 is a summary of BSTS model for predicting international visitor arrival in Indonesia from the US. The smallest posterior mean of the residual standard deviation parameter (residual.sd) is model of Amadeus data. In addition, the smallest standard deviation of the one-step-ahead prediction errors (prediction.sd) is also model of Amadeus data. R-square of Amadeus data model is 0.9803733 which means that the fit explains 98.04 percent of the total variation in the data about the average. Meanwhile, R-square of Model 2 (MPD) shows that the fit can explain 56.42 percent of the total variation in the data about the average and Model 3 (official data) can explain 64.07 percent of the total variation.

Table 5. Correlation between Amadeus Data and Immigration Data, MPD and Immigration Data visitors from USA to Jakarta

Data	Correlation
(1)	(2)
Amadeus Data and Immigration Data	0.5721348
MPD and Immigration Data	0.4699718

The correlation between Amadeus data and official data is shown in Table 5 International visitor arrival from Amadeus data has 0.57 correlation value with official data. Meanwhile, the correlation between MPD and official data accounts for 0.47. The correlation is not high, meaning there is a difference between the immigration data and big data sources. The difference could be caused by the difference in definitions (nationality vs residence) in that travellers with a US passport to Jakarta might not be residing in the US. There are implications for data users, for example the NTO, which is basing its marketing strategy to attract US travellers based on immigration information.

Table 6. Correlation between Amadeus Data and MPD

Data	Correlation
(1)	(2)
Amadeus Data and MPD	0.614607

Prediction of International Visitors from Australia to Jakarta

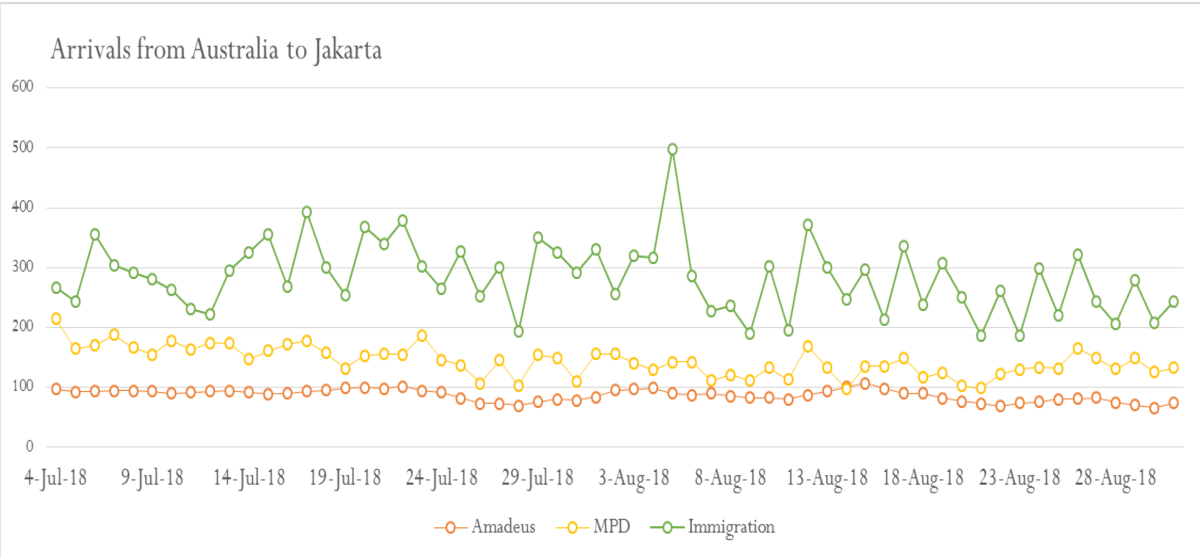

Figure 5. Comparison of arrival data from Immigration, Amadeus and MPD from 5 July to 31 August 2018, Australia to Jakarta

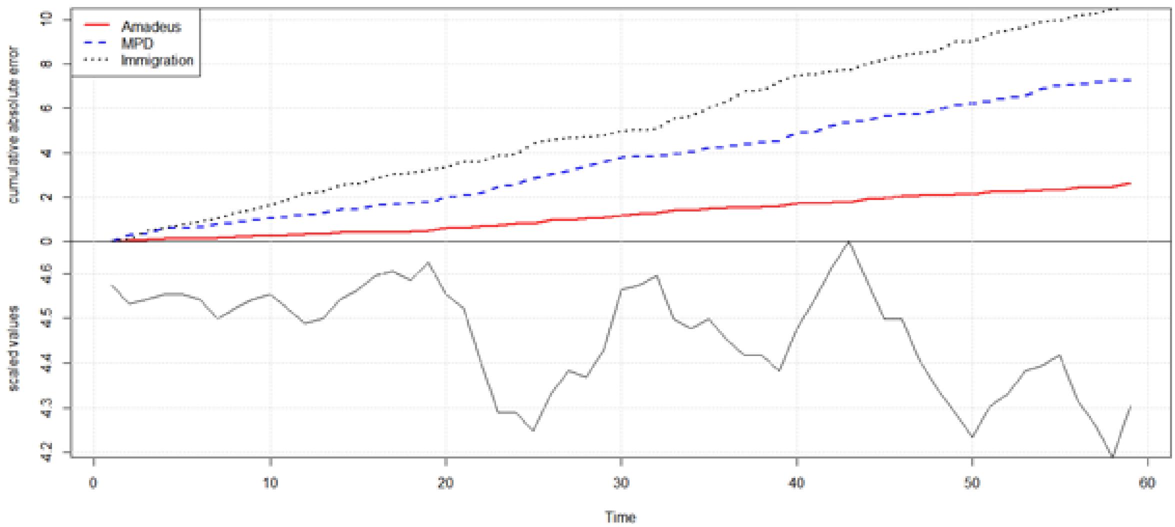

Figure 6. Comparison of Error from Model 1 (Amadeus Data), Model 2 (MPD Data), and Model 3 (Data from Immigration Office) for Australian visitors

Table 6. Summary of Model 1, Model 2, and Model 3 for Australian Visitors to Jakarta

Model	Residual.sd	Prediction.sd	R Square
(1)	(2)	(3)	(4)
Model 1 (Model of Amadeus Data)	0.0236671	0.0615814	0.9585197
Model 2 (Model of MPD)	0.1295624	0.1472879	0.4440538
Model 3 (Model of Immigration Data)	0.1968329	0.2150374	0.1041635

Figure 4 is a summary of BSTS model for predicting international visitor arrival in Indonesia from Australia. The smallest posterior mean of the residual standard deviation parameter (residual.sd) is model of Amadeus data. In addition, the smallest standard deviation of the one-step-ahead prediction errors (prediction.sd) is also model of Amadeus data. R-square of Amadeus data model is 0.9585197 which means that the fit explains 95.85 percent of the total variation in the data about the average. Meanwhile, R-square of Model 2 (MPD) shows that the fit can explain 44.41 percent of the total variation in the data about the average and Model 3 (official data) can explain 10.42 percent of the total variation.

Table 3. Correlation between Amadeus Data and Official Data for Australian Visitors, MPD and ImmigrationData

Data	Correlation
(1)	(2)
Amadeus Data and Immigrationl Data	0.3428086
MPD and Immigration Data	0.4193825

Table 7. Correlation between Amadeus Data and MPD

Data	Correlation
(1)	(2)
Amadeus Data and MPD	0.4431798

The correlation between Amadeus data and official data is low, as shown in Table 3. International visitor arrival from Amadeus data has 0.34 correlation value with official data. Meanwhile, the correlation between MPD and official data accounts for 0.42. Similarly to US arrivals, there is a difference in the definitions on who is included in "Australian arrivals" between immigration and the big data sources. From Australia, there is a high contingent of Indonesian passport holders, who travel from Australia and with Australian SIM cards to Jakarta, which is not shown in immigration data.

6. Conclusions

This paper studies the demand of International visitors from France, United States, and Australia to Jakarta, using new data sources from big data, Amadeus data and Mobile Positioning Data (MPD). The results suggest that both, Amadeus data and MPD, can be used to predict International visitors to Indonesia for some countries, with a good correlation and high goodness of fit. However, for other countries, there is a misalignment in definitions of country of origin.

Data users need to be aware that many tourism marketing organisations are currently using nationality stated in the passport without acknowledging where the tourist lives (country of residence). For example a tourist with Australian nationality is recorded by Immigration data, however, the tourist's residence may not be Australia, it may in fact be Thailand. Amadeus data and MPD captures the country of residence enabling tourism marketing to be more accurate with their expenditure on marketing activity.

The empirical evidence showed that Amadeus and MPD can be used for projections and Amadeus can be used for forecasts/predictions based on forward bookings for the next 12 months, improving the quality of predictive analytics. What is more, all three data sources have advantages. All the three data sources should be used together to support each other limitation. The gate of entry gives the common point of comparison, the benefits of each data source are divided from that: Amadeus for forecasting up to 12 months; Immigration for benchmarking arrivals only; while MPD to see final destinations and movement in the country which is important for tourism destination analysis

Recommend that the Ministry of Tourism utilise Amadeus data and MPD to improve marketing of Indonesia by aligning marketing activity with target origin markets, as this will result in more accurate and more effective policy.

7. References

1. Brodersen, Kay H., et.al. (2015), Infering Causal Impact Using Bayesian Structural Time-Series Models, The Annals of Applied Statistics, Vol. 9, No. 1, pp. 247-274, Institute of Mathematical Statistics 2015.
2. BPS. (2017). Emerging Challenges in Data Collection, Survey Methodology and Implications for Official Statistics (NSO Perspectives), presented in the Plenary Session, International Statistical Institute Regional Statistics Conference (ISI RSC) 2017, Bali.
3. Congdon, Peter (2006), Bayesian Statistical Modelling [Second Edition], London: John Wiley & Sons, Ltd.
4. Landefeld, Steve. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues, Discussion Paper, accessed on 16 March 2017, https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf.
5. Morales et.al. (2014). A World that Counts: Mobilising the Data Revolution for Sustainable Development. The United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG).
6. Pole, Andy, Mike West, and Jeff Harrison (1994), Applied Bayesian Forecasting and Time Series Analysis, US: Springer Science and Business Media, B.V.
7. Samuelson, P.A (1976), Economics. Mc Graw-Hill, New York.
8. Seynaeve, Gerdy and Christophe Demunter. (2016). When Mobile Network Operators and Statistical Offices Meet-Integrating Mobile Positioning Data into the Production Process of Tourism Statistics, *the 14th Global Forum on Tourism Statistics*.
9. Sniowski, Alessandro, Sorichetta, Ingmar Weber, and Andrew J. Tatem (2016). Inferring Migrations: Traditional Methods and New Approaches based on Mobile Phone, Social Media, and other Big Data. *Technical Report*. European Union.
10. WTTC (2019). Travel and Tourism Economic Impact: World, WTTC Report 2019