Novel Approach in Outbound Tourism Statistics in the Era of Revolution Industry 4.0
(Case Study of Indonesia Outbound Tourism Statistics)

Dini Arifatin\textsuperscript{1,a}, Rifa Rafiadi\textsuperscript{1,b}, Ignatius Aditya Setyadi\textsuperscript{1,c}, Ricky Yordani\textsuperscript{2,d}

\textsuperscript{1} BPS-Statistics Indonesia
\textsuperscript{2} Polytechnic of Statistics STIS, Jakarta-Indonesia
Email: \textsuperscript{a}arifatin@bps.go.id; \textsuperscript{b}rifa@bps.go.id; \textsuperscript{c}adityasetyadi@bps.go.id; \textsuperscript{d}yordani@stis.ac.id

Abstract:
The amount of outbound tourism is related to strong economic performance and increment in citizens income (World Tourism Organization, 2019). Therefore, the outbound tourism statistics as an important indicator for measuring the national economic performance. UN Statistics Division (UNSD) recommended for determining the flows of outbound visitors using an entry/departure card; a specific survey at the border, or observing them from household surveys. Previously, Indonesia used the monthly report from Directorate General of Immigration (DGI) collected at the departure gate. The problem was this method cannot distinguish between outbound tourists, nationals residing abroad, and frequent border crossers. The study aimed to develop novel approach that is able to distinguish outbound tourists, nationals residing abroad, and frequent border crossers from the visitors. We proposed novel method on the arrival and departure gate approach were the data obtained daily by DGI using passport scanning machine. This approach paired information from the departure and arrival gate for obtaining the total of length of stays (LOS) in Indonesia and the total trips of visitors. Using the data from 2018 to 2019, we obtained for 2019 the outbound visitors, national residing abroad, and frequent border-crossers were 8,184,361; 664,645 and 117,936 persons, respectively.

Keywords: tourism length of stay; nationals residing abroad; frequent border crossers; immigration data; administrative data

1. Introduction
Outbound tourism refers to activities of visitor crossing the international border from the country of residence. These activities have an impact on the economic directly or indirectly to the destination country as well as a sign of the economic situation in the country of visitor residence. According to World Tourism Organization (2019), the fast rate growth of outbound tourism of China citizens during the last decades driven by strong country economic performance and increasing disposable income of the China citizens. Afterwards, it makes it beneficial to conduct an in-depth study in Indonesia on statistical calculation of outbound tourism.

Outbound activities have criteria whether a trip taken qualifies as a tourism trip. The criteria had to meet the requirement in terms of duration and main purpose of the trip. The duration had to take for less than a year start from the visitor departure to arrival. In addition, the main purpose of trip is one as follows: leisure, business, health, visiting relatives, religion/pilgrimages, shopping, education, meeting or congress, sport, etc., except for working (UN, 2010). Furthermore, in case of the duration trip has more than a year then it is should be considered as international visitors or national residing abroad and in Bahasa named as Penduduk Luar Negeri (pendul). Moreover, for the countries having a land border there is an important movement of persons over the border. This movement generally had many trips during a year. The persons known as frequent border-crossers. UN (2010) declared for frequent border-crosser might present theoretical and practical difficulties on tourism activity as the fact that the population living on the border is often exempted from interacting with immigration authorities.
UNSD recommended for determining the flows of outbound visitors using an entry/departure card; a specific survey at the border, or observing them from household surveys. These techniques also considered for identifying and obtaining the number of outbound visitors, national residing abroad and frequent border-crossers. In 2014, the DGI authority stopped collecting this card. Nevertheless, BPS is still conducted outbound surveys for obtaining administrative data at several arrival gate but not annually. In addition, BPS used calculated ratio based on survey in 2008 and renewed in 2017 for distinguish between national residing abroad or not (BPS, 2017). This method still has problems for distinguishing between the number of outbound visitors, national residing abroad and frequent border-crossers.

DGI authority started to implement of industry 4.0 for recording and digitizing immigration data. In 2018, the monthly data administration report of inbound and outbound visitors from DGI were sent through email to BPS. Afterwards, DGI and BPS collaborated using a special access point based on API (Application Programming Interface) through web service for sharing the immigration data on daily since 2019. It produced tons of records, the analysis of these administrative data presents new statistical challenges and requires Big Data technique to obtain the result.

Immigration data as administrative data with numerous arrival-departure transaction records on every gate are having significant advantages as an alternative data sources to obtain tourism statistics. It gives a new technique for BPS to mainly produce inbound tourism statistics, but also deploys an approach to produce outbound, pendul and frequent border-crossers statistics. Hence, this paper aims to presents a detailed explanation in developing the new technique to produce outbound visitors, pendul and frequent border-crossers statistics using numerous arrival-departure transaction records from domestic passport of Indonesia citizen.

2. Methodology

Data and Variables
Since 2018, DGI and BPS have signed The Cooperation Agreement concerning Immigration Data Exchange to Improve the Quality of Tourism Statistics and Transportation Statistics. A data communication network is established between two parties in order to access immigration data daily. To exchange data, exclusive access is provided through the Application Programming Interface (API) and the hashing algorithm is applied to the data provided so that confidentiality and data security are maintained only between parties. It is collected using an integrated passport scanning machine system from immigration checkpoints across Indonesia at 91 airports and 33 harbors. Nevertheless, for land cross border gates with total around of 79 cross border gates, there are only small number cross border gate succeed to use integrated passport scanning machine system. Furthermore, this study used data start from 2018 to 2019 to accommodate the concept of population within one year used in pendul. The variables obtained are as follows: passport number, crossing type (arrival or departure), gate name, type of passport, crossing timestamp, gender and date of birth.

Big Data Approach
Official statistics are a public good that should efficiently meet user needs yet also reduce the burden on respondents, namely through more intensive use of administrative data. As we know, administrative data differ from survey data in terms of purposes and is reflected in its population frame, the unit of observation, the sample size, and the scope of the data. The data generated is huge in volumes that a big data processing approach is needed to produce these statistics, especially for the needs of the official release on a monthly basis with short processing time. In addition, administrative data are known as a distinctive form of big data.
The data provided daily on JavaScript Object Notation (JSON) format so that it needs several processing steps before the data can be processed as a whole. As illustrated in the Figure 1, integrated systems are built to provide this big data processing approach. At first, the web-based download portal application was developed to directly download daily data with output that was converted to the Comma Separated Values (CSV) format Then, the system is designed with an integrated connected procedure.

Robot automation was built with Kofax Kapow™ software to automate data exchange by executing API requests and eventually writes the JSON output files. Kofax Kapow™ is the Robotic Process Automation (RPA) software that allows users to deploy smart robots to mimic human actions and automate a wide range of manual, repetitive tasks while driving continuous improvements. Kofax Kapow™ provides a platform to process data from structured or unstructured databases, files, email systems, websites, and portals. It handles automated extraction and transformation of data from Excel, XML, XLS, PDF, RSS feeds and from APIs based on SOAP, REST, XML, and JSON (Kofax, 2018).

The extracted JSON files then imported into Microsoft SQL Server database, also the process is automated using the robot automation built. The system implements Openrowset built-in function from Microsoft SQL Server which supports bulk operations through a built-in Bulk provider that enables data from a file to be read and returned as a rowset (SQL Server 2017, 2020). Data preparation is carried out during this procedure step, as well as being used as a backup and recovery platform both in database and JSON files form. Backup and recovery are set in case of a loss and setting up systems that allow data recovery due to data loss. Data ingestion then implemented after the following data preparation is complete from Microsoft SQL Server database into Hortonworks Data Platform using Apache Sqoop™. The Apache Hive™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive.

Figure 1. Integrated systems with big data processing approach
Cleansing Data and Classifying of Visitor Types
Cleansing data is the process of detecting and correcting (or removing) corrupt or inaccurate records for identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data (Wu, 2013). This process is to meet the high quality of data criterion as follows: validity, accuracy, precision, completeness, consistency and uniformity. Furthermore, the process of classifying visitor types is needed to adjust to the criteria used by UNWTO. This study used LOS and total trip of visitor for classifying. Afterwards, these two processes combined for distinguishing the visitors.

3. Result
After data extracted from big data processing, millions of immigration data records produced. Furthermore, the cleansing and classifying process started. Although these data produced by automation passport scanning machine, anomalies data sometimes appears. Next, the process is continued by removing anomalies and followed by filtering only Indonesian passport that are processed. The classification process set the time interval two years back from the observation month and paired the passport number from the arrival and departure data to get the LOS in the country. This study also considers other trips that have no pair. Often when visitor is traveling and arrives at the end of the observation period but then does not know when they will depart later. Next, the imputation made for those overlapping trips so that trips have a pair. These imputations conducted for each departure with the arrival at the beginning of the period and each arrival with the departure at the end of the period. The next imputation made to the trip that has departure during the observation period but arrive before the observation period, the imputation arrival made at the start of observation period. The classification process continues counting the frequent of trips. This study used criteria total trips more than 104 in two years as the cut point for the frequent border-crossers. This cut point number comes from the assumption that the frequent border-crossers have the minimum trips on once in a week in every week during the observation time. The visitor with total trips less than 104 will be counted for the LOS in the country. Furthermore, if the calculation of LOS is more than one year then the visitors is indicated as pendul. Next, if the calculation of LOS is less than one year then the visitors is indicated as outbound visitors. The flowchart of the cleansing and classification process is illustrated in Figure 2.

The number of outbound visitors, pendul and frequent border-crossers in 2019 shown monthly in Table 1. It shown that the outbound visitors, national residing abroad, and frequent border-crossers were 8,184,361; 664,645 and 117,936 persons, respectively. Furthermore, from the three types immigration checks point the percentage of visitor type is shown in Figure 3.
Table 1. Number of national residing abroad, frequent border-crossers and outbound visitors in 2019

<table>
<thead>
<tr>
<th>Type of Visitors</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nationals Residing Abroad</td>
<td>54,298</td>
<td>63,128</td>
<td>59,141</td>
<td>44,611</td>
<td>47,812</td>
<td>75,120</td>
<td>47,181</td>
<td>65,769</td>
<td>55,298</td>
<td>48,071</td>
<td>47,408</td>
<td>56,808</td>
<td>664,645</td>
</tr>
<tr>
<td>Frequent Border Crossers</td>
<td>9,677</td>
<td>8,570</td>
<td>6,655</td>
<td>10,992</td>
<td>12,650</td>
<td>11,565</td>
<td>11,472</td>
<td>10,283</td>
<td>10,520</td>
<td>10,241</td>
<td>7,836</td>
<td>7,475</td>
<td>117,936</td>
</tr>
<tr>
<td>Outbound Visitors</td>
<td>864,872</td>
<td>597,753</td>
<td>778,688</td>
<td>610,294</td>
<td>687,507</td>
<td>781,478</td>
<td>634,337</td>
<td>591,148</td>
<td>648,851</td>
<td>567,574</td>
<td>675,373</td>
<td>746,486</td>
<td>8,184,361</td>
</tr>
<tr>
<td>Total</td>
<td>928,847</td>
<td>669,451</td>
<td>844,484</td>
<td>665,897</td>
<td>747,969</td>
<td>868,163</td>
<td>692,990</td>
<td>667,200</td>
<td>714,669</td>
<td>625,886</td>
<td>730,617</td>
<td>810,769</td>
<td>8,966,942</td>
</tr>
</tbody>
</table>

Figure 2. Flowchart of the cleansing and classification process

Figure 3. The percentage of visitor types from three types immigration checks point
4. Discussion, Conclusion and Recommendations

Discussion
This study success to calculate and distinguish three type of Indonesia citizen visitor. Nevertheless, some cross-border gates still have problem on connecting to the system. This problem solved by using email for reporting the immigration administrative data.

The number of three type of visitors related with national tourism balance published by BPS and national financing balance published by Central Bank of Indonesia. This study could improve its calculation which can be combined with the outbound and inbound survey.

The number of frequent border-crossers is often not interested to be discussed. But the results of this study can be a trigger to discuss matters relating to it, such as the methodology and composition of expenses related to the trip. Furthermore, frequent border-crossers has another mechanism beside passport for entering and leaving the country. Caused they could not be detected in the study. Figure 3 proved only on cross border gate did not have frequent border-crossers, it indicate they exempt for dealing with immigration authorities or has another mechanism for entering and leaving the country. Nowadays, BPS has another solving mechanism for this case. BPS use Mobile Position Data (MPD) for improving the flow number of frequent border-crossers.

Conclusion
Even though this study is still proof of concept study but its success distinguishes the three types of Indonesia citizen visitor.

Recommendations
1. The tourism statistical methodology still needs to be improved. One of the solutions can be done by utilizing the technology.
2. This study could be combined with the MPD for improving the accuracy.

References: