Assessing the use of transaction and location based insights derived from Automatic Teller Machines (ATM's) as near real time "sensing" systems of economic shocks

Dharani Dhar Burra
United Nations Global Pulse
Pulse Lab Jakarta, Indonesia
dharani.burra@un.or.id

Sriganesh Lokanathan
United Nations Global Pulse
Pulse Lab Jakarta, Indonesia
sriganesh.lokanathan@un.or.id

Abstract

Big data sources provide a significant opportunity for governments and development stakeholders to "sense" and identify in near real time, economic impacts of shocks on populations at high spatial and temporal resolutions. In this study, we assess the potential of transaction and location based measures obtained from automatic teller machine (ATM) terminals, belonging to a major private sector bank in Indonesia, to "sense" in near real time, the impacts of shocks across income groups. For every customer and separately for years 2014 and 2015, we model the relationship between aggregate measures of cash withdrawals for each year, total inter-terminal distance traversed by the customer for the specific year and reported customer income group. Results reveal that the model was able to predict the corresponding income groups with 80% accuracy, with high precision and recall values in comparison to the baseline model, across both the years. Shapley values further show that the total inter-terminal distance traversed by a customer across multiple ATM terminals in each year had significantly high resolving power. Descriptive statistics and Kruskal-Wallis testing further confirm these observations, and reveal that customers in the lower-middle class income group, have significantly high median values of inter-terminal distances traversed (7.21 Kms for 2014 and 2015) in comparison to high (2.55 Kms and 0.66 Kms for years 2014 and 2015), and low (6.47 Kms for 2014 and 2015) income groups. Although no major shocks were noted in 2014 and 2015, our results show that lower-middle class income group customers, exhibit relatively high mobility in comparison to customers in low and high income groups. Additional work is required to further realise the potential of the sensing capabilities of this data source in the event of a shock, however its ability to provide high-resolution insights on, who, where and by how much is the population impacted by a shock, can facilitate targeted responses.

1 Introduction

High spatial and temporal resolution monitoring of poverty and its dynamics is essential for planning and implementation of development interventions by governments and development agencies. Additionally, the ability to monitor poverty and related dynamics at high resolutions also enables robust monitoring and impact evaluations. Poverty is a complex, multidimensional construct, and is a combination of income, wealth and consumption [1]. Income represents how much a household earns, while wealth and consumption represent asset investment and expenditure patterns.These three components of poverty, and their interactions vary in relation to spatial and temporal contexts [1].

Preprint. Under review.

It is therefore necessary for governments and their partners to monitor these components at high resolutions, specifically in relation to economic shocks.

In Indonesia, poverty measures are collected by Household Income and Expenditure Survey, National Socioeconomic survey (SUSENAS), National Labour Force Survey (SAKERNAS) etc.[2] These surveys are robust, and are consistently used for planning and implementation of economic interventions. However they are limited by their ability to capture poverty and related dynamics at high spatial and temporal resolutions, especially in the context of economic shocks. For instance, SUSENAS is implemented once in every two years across a limited sample of households. Since poverty measures display extreme variability across space and time, and if an economic shock occurs between two successive rounds of SUSENAS, its impacts on populations would either need to be derived from existing surveys, or one of the surveys would need to be implemented afresh.

Big (or Alternative) data sources provide an alternative and an interesting opportunity for use as early warning and "sensing" systems, to assess impacts across space in near real time. In fact, such sensing capabilities of big data sources have been previously used, and continue to be used in the context of socio-economic development and monitoring in Indonesia. For instance, social media platform Twitter is used by government of Indonesia to "nowcast" market prices of key agricultural commodities, to obtain prices of key food products in near-real time, an insight that wouldn't have been available if development practitioners relied only on traditional administrative data sources[3].

However since big data sources are not designed to sense or monitor development related indicators, a significant amount of preparatory analysis, in terms of what specific aspects of the end user do they capture and how does this translate into insights, is required to assess the applicability of big data sources especially in the context of shocks.

Location and transaction based information derived from ATM (Automatic Teller Machines) terminals offer a potentially interesting opportunity to sense poverty dynamics, at high spatial and temporal resolutions. Economic shocks resulting from events such as the COVID-19 pandemic have significantly large impacts, especially on vulnerable sections of the populations. Furthermore, household financial vulnerability in Indonesia has been shown to be significantly associated with household incomes [4]. However, as none of the traditional surveys provide near real time insights on questions such as who is being impacted, where is the largest impact and by how much, there is an urgent need to develop and test near real time methods, that can sense impacts of shocks across income groups.

Seventy-five percent of the Indonesian population currently have access to banking services, this data set therefore offers a potentially interesting opportunity to "sense" impacts resulting from shocks both across space and time. In this study we report the results from an initial assessment of location and transactions based insights obtained from the ATM network of a major private sector bank in Indonesia, for the years 2014 and 2015. We perform preparatory analysis and assess the feasibility of the data to sense differential impacts of shock across income groups at high spatial and temporal resolutions.

2 Data and Methods

The anonymised data set consists of individual customer transaction logs for years 2014 and 2015, performed at ATM terminals across the island of Java, by bank account holders of a major private sector bank in Indonesia. For each year, every log consists of an anonymised user identification number (ID) that is related to the unique bank account number, type (Cash Withdrawal or Others), and amount associated with each transaction. Additionally, yearly reported incomes for the corresponding unique identification number, as captured in the bank's Know Your Customer (KYC) records was also obtained. In the KYC records, yearly incomes are classified according to the following groups - sd. 1 Juta, >1 Juta - 3 Juta, >3 Juta - 5 Juta, >5 Juta - 10 Juta, >10 Juta - 25 Juta, >25 Juta - 50 Juta, >50 Juta - 100 Juta and >100 Juta, wherein Juta refers to a million Indonesian Rupiah (IDR). Since the proportion of customers within the income groups >3 Juta - 5 Juta (lower-middle class income group; 158358 and 174848 unique customer IDs for years 2014 and 2015), sd. 1 Juta (lower class income group; 33918 and 31109 unique customer IDs for years 2014 and 2015) and >100 Juta (high income group; 11362 and 15398 unique customer IDs for years 2014 and 2015) groups, was greater than 10% across the entire customer base, the data set therefore for each year was subset to these specific income groups.

The data set also contains a geotag for every ATM terminal. The geotag for each ATM terminal is derived from the corresponding nearest postal code. Eighty percent of the transaction logs captured across both the years were tagged as "Cash Withdrawals" as transaction type, therefore logs with this transaction type were further filtered, and used for subsequent analysis. A total of 9 million transaction logs for each year were obtained after this step. Since a customer transacts multiple times across multiple terminals all across the year, aggregate statistics of transactions across multiple cash withdrawals from various terminals, for each customer were derived separately for each year, and used as predictors. Specifically, for each customer, all cash withdrawal transactions across multiple ATM terminals for each year were aggregated, to calculate per year sum, average, standard deviation and median values of cash withdrawal amounts. Also since customers withdraw from multiple ATM terminals, total inter-terminal distance was calculated using the geotags of ATM terminals used by the customer. Per customer, per year aggregate statistics of cash withdrawals, and total inter-terminal distance traversed across ATM terminals were calculated for approximately 260000 customers.

Preprocessing and derivation of predictors was implemented in R statistical environment (v 3.4.1) [5], using the dplyr [6], tidyr [7] and data.table [8] packages. Calculation of the total inter-terminal distance traversed, i e. total distance traversed by individual customer between multiple ATM terminals in a year, was calculated using the Euclidean distance measure (converted to kilometers) using the sp package [9] in R statistical environment (v 3.4.1) [5]. Per customer and per year summary statistics of cash withdrawals (i.e. sum, median, standard deviation, average and total distance traversed) for the selected income groups (i.e. sd 1 juta or lower income group, >3 juta - 5 juta or lower-middle income group and >100 juta or high income group) was calculated using the dplyr package [6] in R statistical environment (v 3.4.1) [5], and the resulting box plots were generated using the ggplot2 [10] package in R statistical environment (v 3.4.1) [5].

For each customer across each year, cash withdrawal aggregates, and total inter-terminal distance traversed for that specific year were used as predictors, the reported income group for the corresponding customer was used as the predictand, and the relationship between the two was modelled in the following steps. The relationship between the predictors and predictand was modelled using Extreme Gradient Boosting (XGB) algorithm, using package xgboost [11] in R statistical environment (v 3.4.1) [5]. XGB is a boosting algorithm that is based on ensemble learning (the final model is based on the collective output of individual weak models). As the predictive power of each individual model is weak, ensembling many weak models provides higher predictive power to the final model. XGB models are relatively easy to tune for hyperparameters, versatile (can be used for binary and multiclass classification with no gain in error) and scalable. Hyperparameter tuning for XGB algorithm was performed using the mlr package [12] in R statistical environment (v 3.4.1) [5]. Tunable parameters which included the booster type, maximum depth, gamma, minimum child weight, subsample, colsample by tree, eta, error metric and evaluation metric, were tuned using a bracket of lower and upper limit values for each parameter. Package CARET [13] in R statistical environment (v 3.4.1) [5] was used to partition the data for training, validation and testing, construction of the confusion matrix, and to obtain the corresponding accuracy, specificity and sensitivity values. For comparison, a baseline model that was tuned to predict the most frequent customer income group, i.e. the lower-middle income group, was developed using the ZeroR function in OneR package [14] in R statistical environment (v 3.4.1) [5]. Accuracy, recall, precision and F1 scores of the XGB model were compared with that of the baseline model, to provide contextual understanding of the results. In order to aid interpretation of the XGB model, variable importance plots using package vip [15], and partial dependence plots were generated using the pdp package [16] in R statistical environment (v 3.4.1) [5]. The predictors used in the models are known to be highly correlated, and since feature importance plots are unable to capture inherent correlations between the features, feature importance plots were produced based on SHAP values and their interactions, using the SHAPforxgboost package [17] in R statistical environment (v 3.4.1) [5]. To further validate the findings of the model, a Kruskal-Wallis hypothesis test was performed to test for significant relationship between the top predictor obtained from the XGB model, and customer income group in R statistical environment (v 3.4.1) [5].

3 Results and Discussion

Descriptive statistics for 2014 and 2015 revealed that high income group (i.e. >100 Juta) had relatively higher mean values of average and median amounts per cash withdrawals in comparison to the lower and lower-middle income groups, i.e. sd. 1 Juta and >3 Juta - 5 Juta, in 2014 and 2015 as seen in figure 1. Additionally figure 1 shows that yearly mean values of standard deviation per cash withdrawal, and the total inter-terminal distance traversed across multiple ATM terminals by the high income group was lower in comparison to the low and low-middle income group in 2014 and 2015.

Predictive modelling using XGB, wherein customer's income group (obtained from the banks KYC records), was predicted using predictors derived from cash withdrawal transactions at the ATM terminal, i.e. per customer and per year sum, average, median, standard deviation values across all cash withdrawal transactions, and per year per customer total inter-terminal distance traversed by the respective customer, revealed that the customer income group can be predicted with 80% accuracy, with relatively high recall values (0.53 for high income group, 0.79 for lower-middle and 0.59 for low income group for 2014/0.59 for high, 0.80 for lower-middle and 0.60 for low income group for 2015). However, the model yielded relatively lower precision values for both the years (0.13 for high, 0.97 for lower-middle and 0.14 for low income group for 2014/0.10 for high, 0.98 for lower-middle and 0.05 for low income group for 2015). In contrast, the baseline model obtained a predictive accuracy of 78% with null values for precision and recall for low and high income groups, for both the years. The XGB model therefore provided significantly better results in terms of predictive power in comparison to the baseline model. As seen in figures 2 and 3,feature importance plots based on SHAP values for 2014 and 2015, obtained for each predictor, revealed that the per year total inter-terminal distance traversed across multiple ATM terminals by a customer was the most important predictor across both the years. SHAP values further suggest that the resolving capabilities of total inter-terminal distance traversed across multiple ATM terminals is significantly higher for lower income groups (blue colored dots for this predictor to the left in in figures 2 and 3). Partial dependence plots as seen in figure 4, for total inter-terminal distance traversed across multiple ATM terminals in 2014 and 2015, further confirms that the probability of predicting lower-middle class income group using the total per year per customer inter-terminal distance traversed, is significantly higher than low and high income groups. Kruskal-Wallis test further confirmed this finding, as the test identified a significant difference in total inter-terminal distance traversed per year between the customer income groups (Kruskal Wallis Chi-squared test, p-value < 2.2e-16), with significantly high median values for low-middle income group (7.21 kilometers in 2014 and 2015) in comparison to high (2.55 in 2014 and 0.67 kilometers in 2015) and low income group (6.47 kilometers in 2014 and 2015).

These results reveal that customers in the lower-middle income group perform cash withdrawals across diverse ATM terminals consistently, in comparison to low and high income group customers, suggesting that customers in the lower-middle income group are relatively more mobile. In fact it is known that Indonesia's lower-middle class does display relatively higher mobility, specifically in relation to income generation opportunities [18], which is reflected by these results.

There were no notable economic shocks in 2014 and 2015, which is also reflected by the fact that lower-middle income group show higher mobility in comparison to low and high income group. However, one can potentially extrapolate the findings to a shock event like the COVID-19 and related government interventions. For instance, relatively lower inter-terminal distances traversed specifically by the lower-middle income group, in comparison to the counterfactual predicted values, can potentially reflect the impacts of shock. However to achieve this, further work is needed in terms of analysis of ATM terminal transaction logs with higher temporal frequencies and spatial resolutions, along with improved modelling approaches, such as use of neural networks and deep learning approaches, and development of relevant insight validation methods etc. Although, there is potential for "sensing" shocks and its impacts in near real time, further work is needed to realise the full potential of this data, in order to facilitate its use for design and implementation of timely responses to shocks by governments and development organisations.

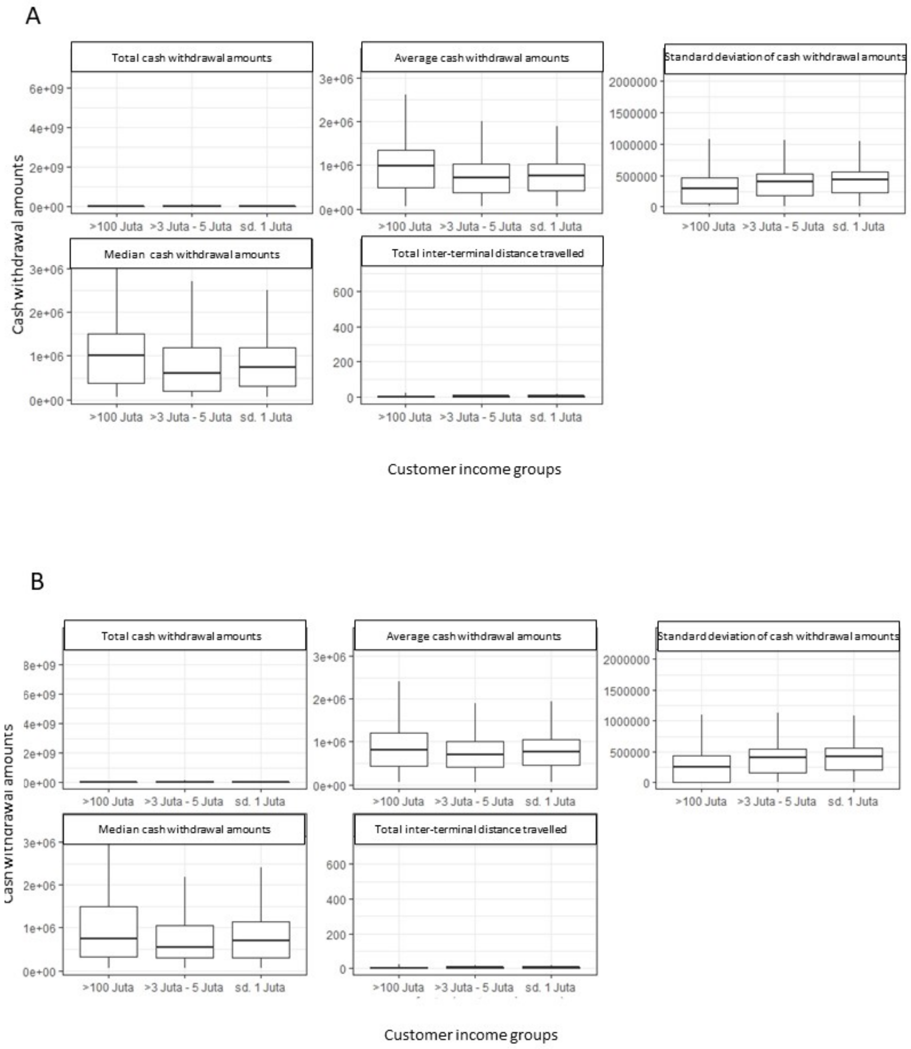

Figure 1: Descriptive Statistics of transaction and location based characteristics obtained from Automatic teller machines (ATM's) for (A) 2014 and (B) 2015

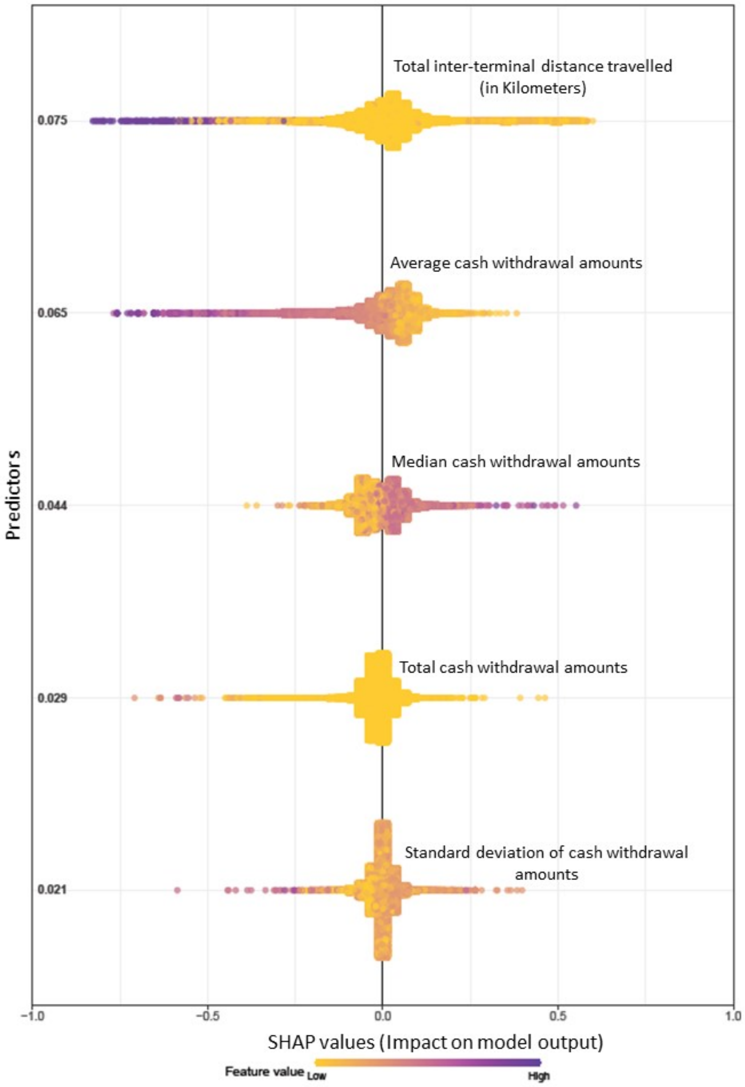

Figure 2: Variable importance plots reflecting the SHAP values for predictors used in the Extreme gradient boosting model in 2014

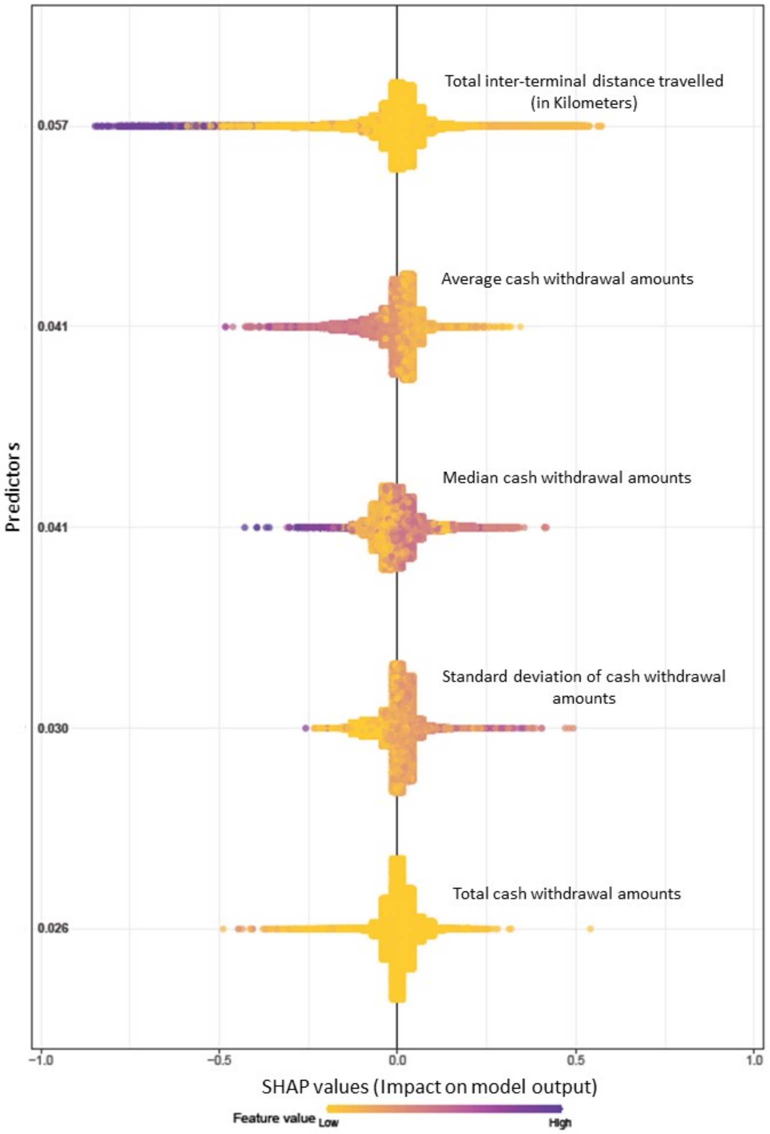

Figure 3: Variable importance plots reflecting the SHAP values for predictors used in the Extreme gradient boosting model in 2015

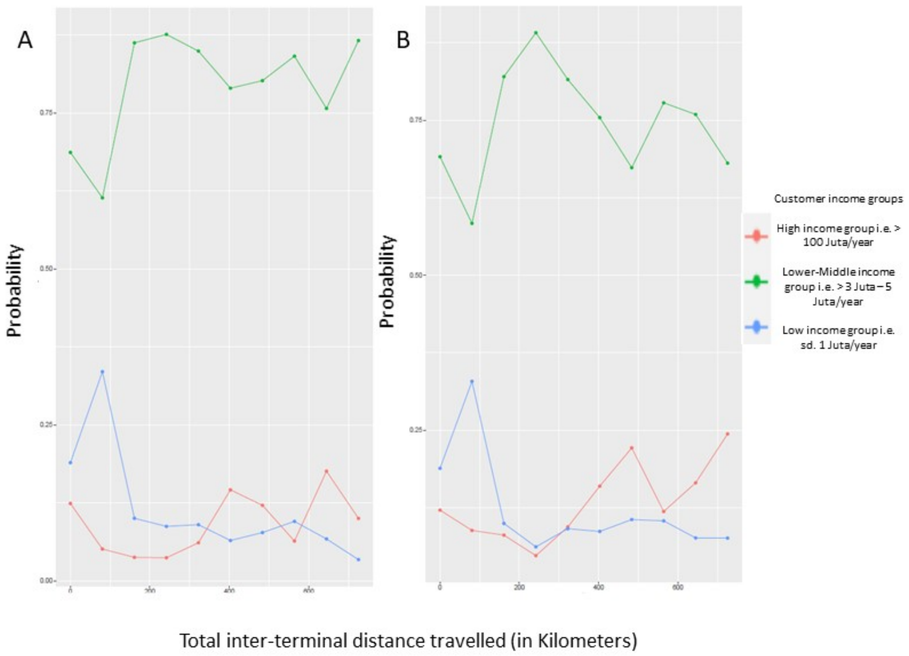

Figure 4: Partial dependence plot of the total inter-terminal distance traversed by a customer for cash withdrawals in (A) 2014 and (B) 2015

Broader Impact

In this work we present evidence in relation to the opportunity offered by transaction and location based insights obtained from Automatic Teller Machines (ATM's), to sense economic shocks and their impacts on populations in near real time, to specifically understand who is being impacted, where is the impact,and by how much. Although these insights are helpful in enabling governments and development agencies to respond to such shocks, it relies on a data source that is confidential and not collected/maintained by either the government or development agencies. This data belongs to the banks and credit rating agencies, and is used to understand the financial characteristics,and spending patterns of their customers. Banks and credit rating agencies are often interested in longer term effects of shocks on their customers, while governments and their partners are interested in relatively short to medium term impacts. Therefore it is necessary to balance the interests of the banks, and in response gain access to their data resources. For instance, new use cases from this data set, such as the relationship between customers' response to short term impacts to the long term ones, and the resulting changes in characteristics such as credit-worthiness, that can potentially help the bank should be explored further. One particular challenge currently with the data source is customer incomes. Different banks have varying capacities in validating customer reported incomes, which will significantly impact the quality of the data and derived insights. Additionally there is an intriguing spatial bias in the methodology too, as we currently assume that customers transacting at a specific terminal also reside in the same location, especially since we don't have access to the customers home location. In conclusion, this is the first time wherein ATM transaction data is being explored for use as a real time economic-shock sensing system in Indonesia, although the opportunities are galore in terms of enabling rapid response by government agencies and their partners, additional exploration will potentially reveal further shortcomings of the data, methods and the corresponding assumptions.

Acknowledgments and Disclosure of Funding

The authors thank colleagues at Bank OCBC-NISP for their support in enabling access to transactions logs from their Automatic Teller Machines. The authors would also like to thank the contributions of team members at Pulse Lab Jakarta, specifically in relation to reviewing this work. Authors would like to thank the Department of Foreign Affairs and Trade of the Government of Australia, along with Ministry of National Development Planning (BAPPENAS) of the Government of Indonesia for providing consistent support and funding for this work. The authors would like to declare that no competing interests exist for this work.

References

[1] Pierre Lamarche, Friderike Oehler, and Irene Rioboo. *European household's income, consumption and wealth*. Statistical Journal of the IAOS Preprint: 1-14.

[2] Robert Genthner and Krisztina Kis-Katos. "Local labor market effects of FDI regulation in Indonesia". 2019.

[3] Imaduddin Amin et al. "Social media insights for sustainable development and humanitarian action in Indonesia". In: *JPhCS* 971 (2018), p. 1.

[4] Sri Noerhidajati et al. *Household financial vulnerability in Indonesia: Measurement and determinants*. Economic Modelling, 2020.

[5] R. Core Team. *R: A language and environment for statistical computing*. Vienna, Austria. URL: R Foundation for Statistical Computing, 2013. URL: http://www.R-project.org/.

[6] Hadley Wickham et al. "dplyr: a grammar of data manipulation, 2013". In: *URL com/hadley/dplyr. version 0.* 1 (2017). URL: https://github.

[7] Hadley Wickham and Lionel Henry. "tidyr: Tidy Messy Data". In: *R package version* 1 (2020), p. 1. URL: https://CRAN.R-project.org/package=tidyr.

[8] Matt Dowle and Arun Srinivasan. "data.table: Extension of 'data.frame'". In: *R package version* 1.12 (2019), p. 8. URL: https://CRAN.R-project.org/package=data.table.

[9] E. J. Pebesma and R. S. Bivand. "Classes and methods for spatial data in R". In: *R News* 5 (2005), p. 2. URL: https://cran.r-project.org/doc/Rnews/.

[10] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. New York: pringer-Verlag, 2016.

[11] Tianqi Chen et al. “xgboost: Extreme Gradient Boosting”. In: *R package version* 1 (2020). URL: https://CRAN.R-project.org/package=xgboost.

[12] B. Bischl et al. “mlr: Machine Learning in R”. In: $journalofMachineLearningResearch,*$ 17 (2016), pp. 1–5. URL: http://jmlr.org/papers/v17/15-066.html%3E.

[13] Max Kuhn. “caret: Classification and Regression Training”. In: *R package version* 6 (2020). URL: https://CRAN.R-project.org/package=caret.

[14] Holger von Jouanne-Diedrich. “OneR: One Rule Machine Learning Classification Algorithm with Enhancements”. In: *R package version* 2 (2017), p. 2. URL: https://CRAN.R-project.org/package=OneR.

[15] Brandon Greenwell, Brad Boehmke, and Bernie Gray. “vip: Variable Importance Plots”. In: *R package version 0.* 2 (2020), p. 2. URL: https://CRAN.R-project.org/package=vip.

[16] Brandon M. Greenwell. “pdp: An R Package for Constructing Partial Dependence Plots”. In: *The R Journal* 9.1 (2017), pp. 421–436. URL: https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

[17] Yang Liu and Allan Just. “SHAPforxgboost: SHAP Plots for ’XGBoost’”. In: *R package version 0.0* 4 (2020). URL: https://CRAN.R-project.org/package=SHAPforxgboost.

[18] Iwan Rudiarto, Rizqa Hidayani, and Micah Fisher. “The bilocal migrant: Economic drivers of mobility across the rural-urban interface in Central Java, Indonesia”. In: *Journal of Rural Studies* 74 (2020), pp. 96–110.