ANALYZING PODES DATA

Research Report for Faculty of IT, Monash University

Yu (Robert) Fu

Master of Data Science

Monash University

CONTENTS:

Part I. Background and Introduction

[Key Research Question]

In recent years, the Indonesian government has been carrying out major expansions of national grid (PLN) to provide rural villages with access to electricity. The pace and magnitude of each year's expansion are believed to be characterized by village economic and geographical feasibilities. It is of interest, therefore, to identify the "disadvantaged" villages (those unlikely to be connected to PLN soon) and advise with viable alternative power options.

This report details the procedures and results from the initial analysis of multi-year Indonesian Village Surveys (PODES) in understanding factors that identify these "disadvantaged" villages. While working as a winter and on-going part-time research assistant at Faculty of Information Technology at Monash University, I was luckily supervised by Dr. Christoph Bergmeir and Professor Lachlan Andrew, as long as under general guidance of Professor David Taniar and Ariel Liebman, to whom I owe my great thanks for providing me with such a learning experience.

My primary research questions are:

a. From PODES data, find out what differs in descriptive statistics between PLN-connected villages and those unconnected.
b. From analyzing survey responses of 2011 and 2014, predict village PLN connect status after

2014.

In order to acquire a certain level of interpretability, I was suggested to use a simple prediction model. In addition, for future simulations, it is recommended that model results be presented with probabilities.

[Main Tasks]

To answer the research questions, the first and foremost step is to record the change of village connection status in a given year. In particular, we would like to know among all disconnected villages in one survey (for instance, 2011), which would change connection status by the next survey (say, 2014). This creates a group of "***connect***" villages that differ from the rest of disconnected ones in a given year, and thus form the basis of our comparison and analysis.

There are three major tasks involved in the analysis that will be detailed in the report. The first part is running descriptive statistics between "***connect***" villages with others - namely, still-yet-to-connected villages, and already-connected ones. The second part tries to create a uniform schema for 2011 and 2014 surveys, and use that for training the predictive model. The third task is to report major findings and discuss for future steps.

[Environment and Packages]

All coding is done using R 3.31 in Jupyter Notebook for easy presentation. Packages used include:
XLConnect – for excel file input
dplyr – for data manipulation
Fselector – for calculating information gain
Glmnet – for penalized regressions
AUC – for calculating AUC and ROC
doMC – for faster parallel caculation.
H2O – for neural network setup

Part II. Datasets Description and Pre-processing

[Datasets]

Using the provided PODES survey from 2008, 2011 and 2014 as the basis of my research, the table below summarizes the number of observations (villages) and questions (variables) in each dataset. A chronological comparison of village IDs shows after each survey, around 90 percent of villages would remain till the next survey, while new villages would comprise about 10 percent in total. This is understandable due to possible administrative causes in real life, and therefore, for the interest of analysis, only villages belonging to both consecutive years (2008-2011, or 2011-2014) are used.

Dataset	**# of Villages**	**# of Variables**	**# of Villages Remained from Last Survey**	**# of New Villages**
2008	75,410	560	NA	NA
2011	77,961	580	69,873 (93%)	8,088 (10%)
2014	82,190	592	73,951 (89%)	8,239 (10%)

[Major Challenges]

The original datasets present some challenges in both language and schema, so the following treatments are applied.

a. We used Google Translate to transform the documents from Indonesian to English in order to understand each question. But different wordings may still cause some confusion.
b. Survey schemas are different for PODES 2011 and 2014. Mapping techniques are applied in order to create a uniform template.
c. Some variables are missing from the datasets, or contain very different codings compared with survey documents. After confirming with John Paterson, they are excluded from the predictive models.

[Imputation and Pre-processing]

The ordinal and tiered structure of PODES survey questions introduced tremendous level of missing data for both 2011 and 2014. But they can be treated with a simple imputation method. Except for very few variables (e.g., Age), all others with missing values can be imputed with either “NA” or 0. This is because the missing value is indeed a result of “Not Applicable”, or zero in quantity.

In addition, population of villages is added to the 2014 dataset, using the provided “Podes_Tabular_2014.xlsx” by matching village IDs. For both 2011 and 2014 datasets, “***connect***” variable is created by matching village IDs between PODES data 2008 and 2011, as well as 2011 and 2014, respectively, to acquire the villages that would be connected in the later survey.

Finally, all variable types are checked before importing the dataset to ensure they have the right format.

Part III. Descriptive Statistics

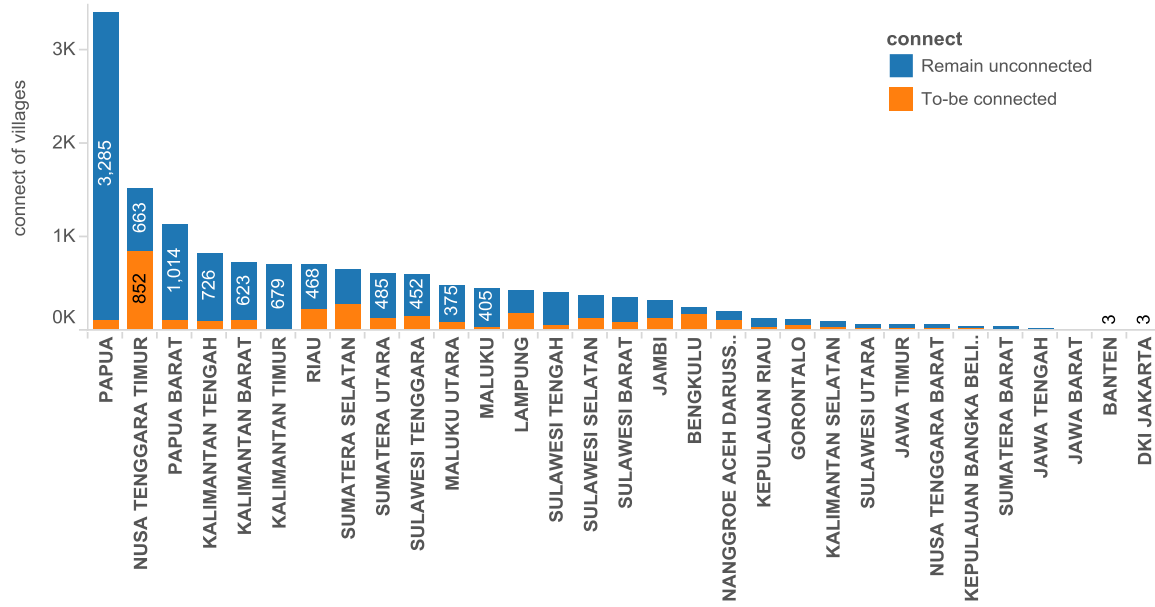
In order to gain an initial idea of how "***connect***" villages are different from the rest of the group, information gain tests are run on both 2008 and 2011 PODES to acquire the independent pair scores between "***connect***" variable with all other variables, and high-ranking pairs are selected for plotting (in Tableau). Though this selection process is rather subjective, it does provide some evidence of potential factors to the outcome variable among the five hundred plus candidates.

The top 100 variables that share highest mutual information with "***connect***" status span across several areas, including: geo-location (i.e., province, district), access to health and educational facilities, road condition and infrastructure, local markets (access to shops, for instance), security conditions (police station or village guards), population characteristics (i.e., size, ethnicity, religion), and revenue from different levels of government. The following section provides descriptive statistics in these highlighted areas between three major village groups of interest, namely:

1) villages that already connected to PLN
2) "***connect***" villages that are not yet but would connected to PLN by the time of next survey
3) villages that would remain unconnected to PLN

Geolocation:

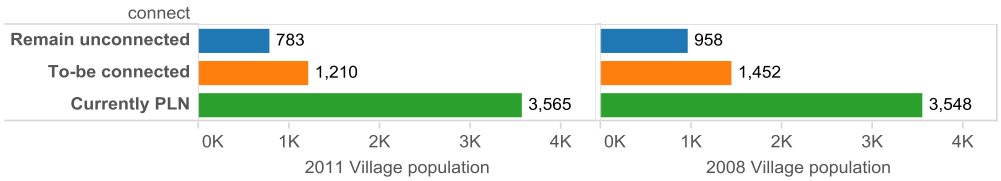
NON-PLN VILLAGES BY PROVINCE IN 2011 PODES SURVEY



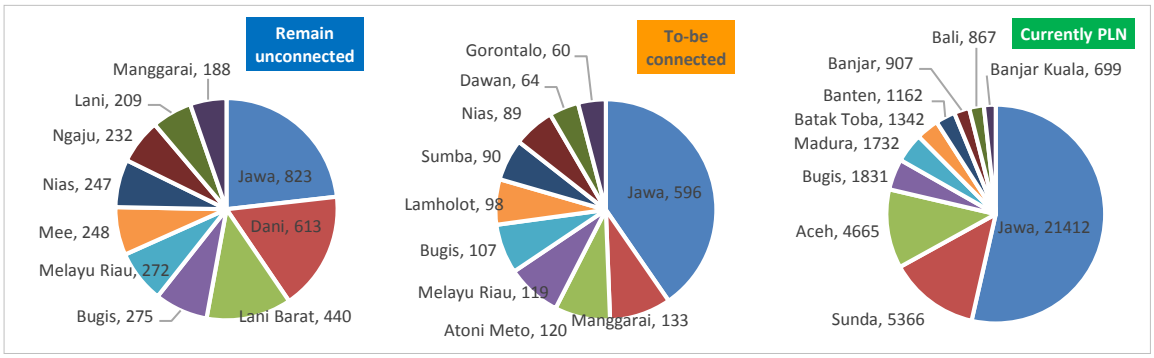
In 2011, Papua has the largest number of villages disconnected to PLN, while Nusa Tenggara Timur has the largest number of "***connect***" villages, showing there would be a great expansion in PLN network between 2011 and 2014. It is believed that province and other geological factors play an important role in PLN connection for the next phase.

Population Characteristics:

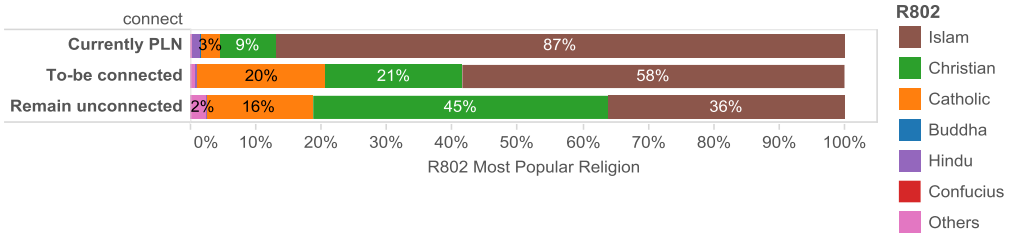
VILLAGE POPULATION SIZE



TOP 10 VILLAGE ETHNICITIES BY CONNECT STATUS IN 2011



VILLAGE PROMINENT RELIGION

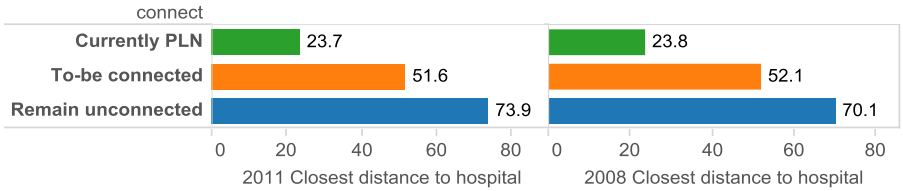


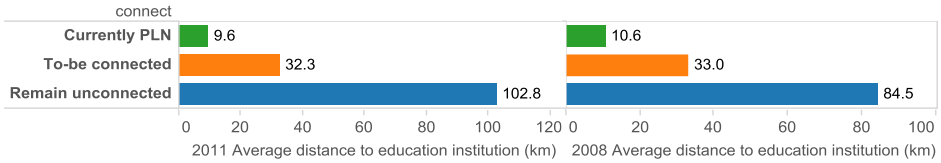
Overall, disconnected villages tend to be smaller in population size, have a greater proportion of ethnic minorities and non-Islamic religions.

Access to Health and Educational Facilities:

It is not surprising to observe the discrepancies of distance to nearby hospitals and education institutions across village types. In general, it seems village connection status are related to the proximity of these facilities.

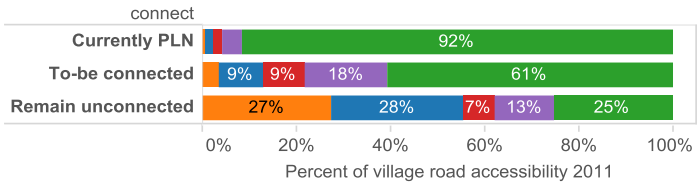
DISTANCE TO HOSPITAL AND EDUCATION INSTITUTIONS

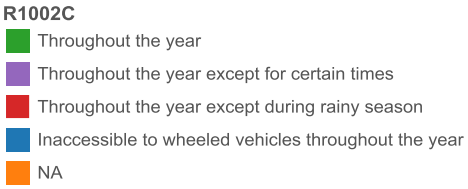


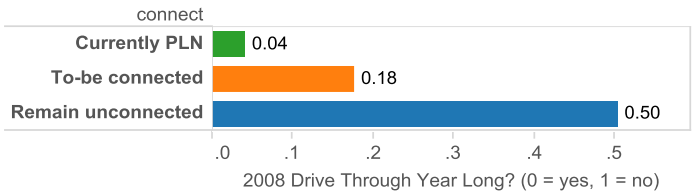


Road Condition and Infrastructure:

Main Road Accessibility

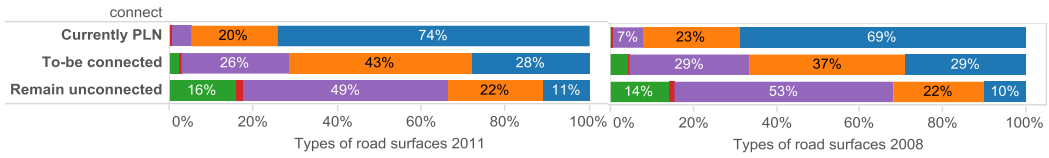


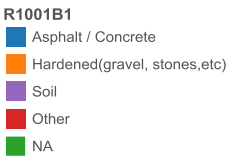




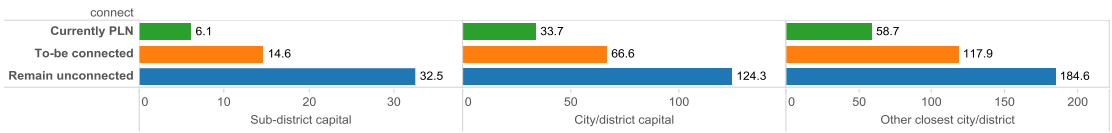
Disconnected villages have less accessible and lower level of infrastructure compared to "***connect***" and PLN-established villages. This could be an indicator that they are located in more remote areas.

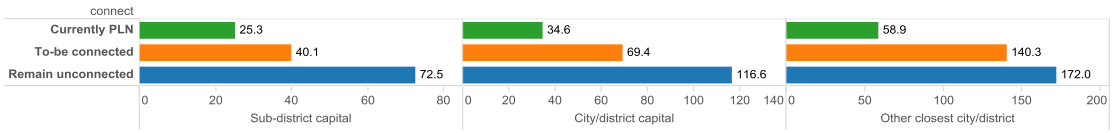
Type of Road Surface





Distance to Nearby Capitals

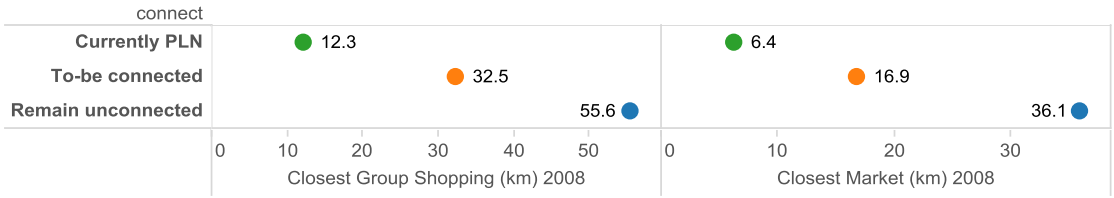


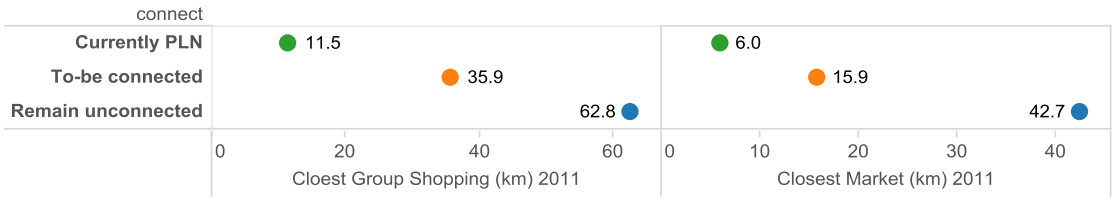


The distance to nearby administrative capitals suggests supporting evidence of their remote location. In particular, the distance between regional capitals and disconnected villages are twice as far as the distance to "***connect***" villages, and four times as far as PLN-connected villages.

Local Markets:

DISTANCE TO MARKETS

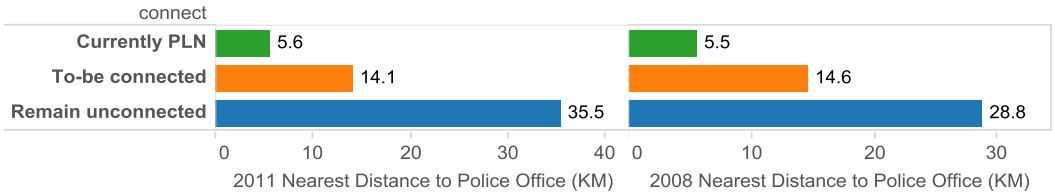




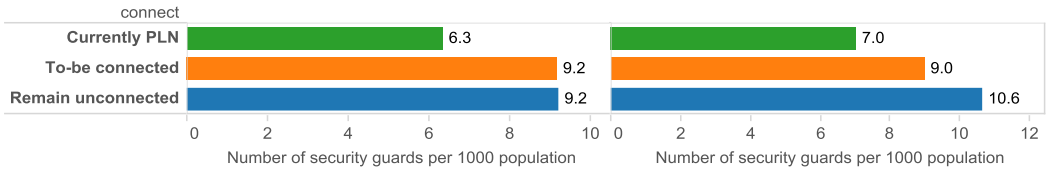
The distance and accessibility to markets and shops suggest the same thing. In general, disconnected villages are twice and four times further away compared the other types of villages.

Security:

DISTANCE TO POLICE OFFICE



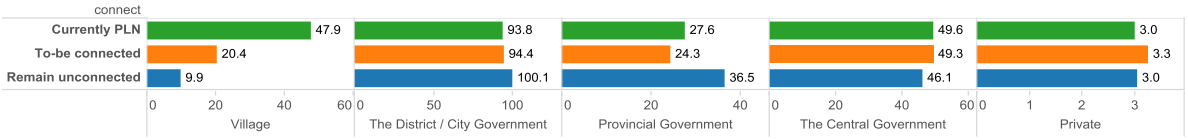
2011 (LEFT) AND 2008 (RIGHT) SECURITY GUARDS RATIO



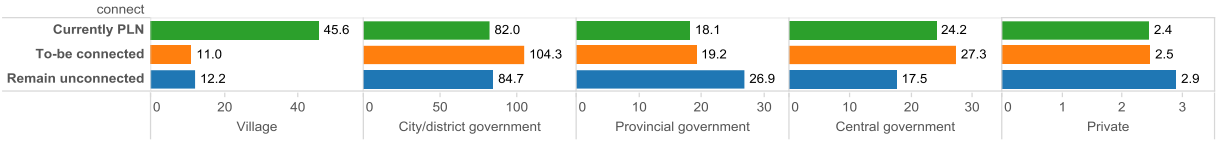
Disconnected villages are also further away from police offices. This might be related to their relatively smaller population sizes and different social context required to keep order. Instead, they have a higher ratio of security guards among population.

Revenue from Different Levels of Government:

2011 VILLAGE REVENUE BY SOURCE (MILLION RP)



2008 VILLAGE REVENUE BY SOURCE (MILLION RP)



There is a drastic difference in village-generated revenue between PLN-established villages and disconnected ones. Although the causal relationship between village connection status and revenue generated or received can be convoluted in nature (it could be very likely that connected villages tend to generate more income), this graph suggests that they might be somehow related to each other. Besides, it is also noted that disconnected villages tend to receive more financial support from the provincial government, but not at sub-district or central level. Considering their smaller population size, this benefit could translate into high number for each person.

To sum up from the descriptive statistics, there seems to be major differences in multiple social and economic aspects between villages that would be connected to PLN, those never connected, and those already connected. In particular, disconnected villages tend to be distributed unevenly across provinces, they have smaller population and greater proportion of religious or ethnical minorities, generally located further away from school, hospital, markets or regional capitals, and have lower grade of transportation and infrastructure. In the end, they seem to be poorer and receive more government support than self-generated incomes.

Part IV. Schema Matching for PODES 2011 and 2014

In order to make sensible predictions for future "***connect***" villages after 2014, a unified schema is required to train and test the mode using 2011 PODES data. While PODES 2011 and 2014 share most of common variables, there are some with different coding or types of answers. Overall, 417 variables from 585 (2011) and 594(2014) are retained in an effort to acquire as much information as possible. It is important to note that the matching process is conducted on PODES 2014 and 2011 after all PLN-connected villages are excluded. The following section reports the major changes to the original dataset by areas of variables.

1. Village Identification

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
villageID	id2011	id2013	Match
provinceID	KODE_PROV	R101	Recode 2014
districtID	KODE_KAB	R102	Match

subDislD	KODE_KEC	R103	Recode 2014

ProvinceID is used instead of Province names because it presents less inconsistency between the two datasets. The one unique province for PODES 2014 “KALIMANTAN UTARA”, which created in 2012 from province “KALIMANTAN TIMUR”, is recoded as the latter to keep consistency. There are also 17 sub-districts corresponding to 108 villages unique in PODES 2014, mostly located in Pupua. Since it requires tremendous domain knowledge to effectively impute them with known sub-districts, a decision is made to delete those villages so that we can train the model properly.

2. General Information of Village

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
villageStatus	R301	R301	Recode 2014
SLS	R304A	R304A	Recode 2011, 2014
SLS_N	R304C	304BA(BCD)_K5	Recode 2014
slope	R305A	305B	Recode 2011
atSea	R305D	307A	Match
atSeaUse_1	R305E2A	307B1A	Match
atSeaUse_2	R305E2B	R307B1B	Match
atSeaUse_3	R305E2C	R307B1C	Match
atSeaUse_4	R305E2D	R307B1D	Match
atSeaUse_5	R305E2E	R307B1E	Match
atSeaMangrove	R305E3	R307B2	Match
byForest	R306A	R308A	Match
byForestUse	R306B	R308B	Match

There are minor coding differences between PODES 2011 and 2014 in village status, the number of SLS facilities, and topography. They are however straightforward and easy for recoding.

3. Population and Employment

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
migrantM	R402A	403B1	Match
migrantF	R402B	403B2	Match
incomeSource	R403A	404A	Match
agriType	R403B	404B1	Match

4. Housing and Environment

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
strLight	R502A	R502A	Match
strLightEngy	R502B	R502B	Match
fuelCook	R503	R503	Match

toilet	R504	R504	Match
litter	R505A	R505A	Match
pubTrash	R505B	R505B	Match
river	R506AK2	R508A_K2	Match
riverUse_1	R506B1K2	R508B1_K2	Match
riverUse_2	R506B2K2	R508B2_K2	Match
riverUse_3	R506B3K2	R508B3_K2	Match
riverUse_4	R506B4K2	R508B4_K2	Match
riverUse_5	R506B5K2	R508B6_K2	Match
irrigation	R506AK3	R508A_K3	Match
lake	R506AK4	R508A_K4	Match
irrigationUse_1	R506B1K3	R508B1_K3	Match
lakeUse_1	R506B1K4	R508B1_K4	Match
irrigationUse_2	R506B2K3	R508B2_K3	Match
lakeUse_2	R506B2K4	R508B2_K4	Match
irrigationUse_3	R506B3K3	R508B3_K3	Match
lakeUse_3	R506B3K4	R508B3_K4	Match
irrigationUse_4	R506B4K3	R508B4_K3	Match
lakeUse_4	R506B4K4	R508B4_K4	Match
lakeUse_5	R506B5K4	R508B6_K4	Match
byRiver	R508A	509B	Match
byRiver_village	R508B	R509C1	Match
byRiver_house	R508C	R509C2	Match
byRiver_family	R508D	R509C3	Match
bySUTET	R509A	R510A	Match
bySUTET_village	R509B1	R510C1	Match
bySUTET_house	R509B2	R510C2	Match
bySUTET_family	R509B3	R510C3	Match
bySlum	R510A	R511A	Match
bySlum_village	R510B1	R511B1	Match
bySlum_house	R510B2	R511B2	Match
bySlum_family	R510B3	R511B3	Match
pollution_1	R511AK2	R512A_K2	Match
pollution_1Src	R511AK3	R512A1_K3	Match
pollution_1Cmplt	R511AK4	R512A_K4	Match
pollution_2	R511BK2	R512B_K2	Match
pollution_2Src	R511BK3	R512B1_K3	Match
pollution_2Cmplt	R511BK4	R512B_K4	Match
pollution_3	R511CK2	R512C_K2	Match
pollution_3Src	R511CK3	R512C1_K3	Match
pollution_3Cmplt	R511CK4	R512C_K4	Match
burnField	R512A	R513	Match
excavation	R513	R514	Match

Despite some wording difference, most questions in this part of the survey are common in both PODES datasets, and are matched in coding.

5. Natural Disaster

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
disaster_1	R60101K2	R601A_K2	Match
disaster_1N	R60101K3	R601A_K3+K5+K7	Recode 2014
disaster_1Dead	R60101K4	R601A_K4+K6+K8	Recode 2014
disaster_2	R60102K2	R601B_K2	Match
disaster_2N	R60102K3	R601B_K3+K5+K7	Recode 2014
disaster_2Dead	R60102K4	R601B_K4+K6+K8	Recode 2014
disaster_3	R60103K2	R601C_K2	Match
disaster_3N	R60103K3	R601C_K3+K5+K7	Recode 2014
disaster_3Dead	R60103K4	R601C_K4+K6+K8	Recode 2014
disaster_4	R60104K2	R601D_K2	Match
disaster_4N	R60104K3	R601D_K3+K5+K7	Recode 2014
disaster_4Dead	R60104K4	R601D_K4+K6+K8	Recode 2014
disaster_5	R60105K2	R601E_K2	Match
disaster_5N	R60105K3	R601E_K3+K5+K7	Recode 2014
disaster_5Dead	R60105K4	R601E_K4+K6+K8	Recode 2014
disaster_6	R60106K2	R601F_K2	Match
disaster_6N	R60106K3	R601F_K3+K5+K7	Recode 2014
disaster_6Dead	R60106K4	R601F_K4+K6+K8	Recode 2014
disaster_7	R60107K2	R601G_K2	Match
disaster_7N	R60107K3	R601G_K3+K5+K7	Recode 2014
disaster_7Dead	R60107K4	R601G_K4+K6+K8	Recode 2014
disaster_8	R60108K2	R601H_K2	Match
disaster_8N	R60108K3	R601H_K3+K5+K7	Recode 2014
disaster_8Dead	R60108K4	R601H_K4+K6+K8	Recode 2014
disaster_9	R60109K2	R601I_K2	Match
disaster_9N	R60109K3	R601I_K3+K5+K7	Recode 2014
disaster_9Dead	R60109K4	R601I_K4+K6+K8	Recode 2014
disaster_10	R60110K2	R601J_K2	Match
disaster_10N	R60110K3	R601J_K3+K5+K7	Recode 2014
disaster_10Dead	R60110K4	R601J_K4+K6+K8	Recode 2014

Both PODES surveys have 10 disaster types. However, in 2014 the number of occurrences and casualties are broken down into three years (2011-2013), while in PODES 2011, they are answered in three-year's total. Therefore, we add columns in PODES 2014 to match with 2011.

6. Education and Health

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
edu_1State	R701AK2	R701A_K2	Match
edu_1Private	R701AK3	R701A_K3	Match
edu_1Distance	R701AK4	R701A_K4	Match
~	~	~	~
edu_5State	R701EK2	R701E_K2	Match
edu_5Private	R701EK3	R701E_K3	Match
edu_5Distance	R701EK4	R701E_K4	Match
edu_6State	R701FK2	R701F_K2	Match
edu_6Private	R701FK3	R701F_K3	Match
edu_7State	R701GK2	R701G_K2	Match
edu_7Private	R701GK3	R701G_K3	Match
edu_8Private	R701HK3	R701H_K3	Match
edu_9Private	R701IK3	R701I_K3	Match
edu_10Private	R701JK3	R701J_K3	Match
voc_1N	R702A	R703A	Match
~	~	~	~
voc_7N	R702G	R703G	Match
otherEdu_1	R703A	R702A	Match
~	~	~	~
otherEdu_5	R703E	R702F	Recode 2014
health_1	R704AK2	R704A_K2	Match
health_1N	R704AK3	R704A_K3	Match
health_1Distance	R704AK4	R704A_K4	Match
health_1EasytoTravel	R704AK5	R704A_K5	Match
~	~	~	Match
health_3	R704CK2	R704F_K2	Match
health_3N	R704CK3	R704F_K3	Match
health_3Distance	R704CK4	R704F_K4	Match
health_3EasytoTravel	R704CK5	R704F_K5	Match
health_4	R704DK2	R704C_K2 R704D_K2	Recode 2014
health_4N	R704DK3	R704C_K3 + R704D_K3	Recode 2014
health_4Distance	R704DK4	min(R704C_K4, R704D_K4)	Recode 2014
health_4EasytoTravel	R704DK5	round(mean(R704C_ K5, R704D_K5)	Recode 2014
health_5	R704EK2	R704E_K2	Match
health_5N	R704EK3	R704E_K3	Match
health_5Distance	R704EK4	R704E_K4	Match
health_5EasytoTravel	R704EK5	R704E_K5	Match
health_6	R704FK2	R704G_K2	Match
health_6N	R704FK3	R704G_K3	Match

health_6Distance	R704FK4	R704G_K4	Match
health_6EasytoTravel	R704FK5	R704G_K5	Match
health_7	R704GK2	R704H_K2	Match
health_7N	R704GK3	R704H_K3	Match
health_7Distance	R704GK4	R704H_K4	Match
health_7EasytoTravel	R704GK5	R704H_K5	Match
health_8	R704HK2	R704I_K2	Match
health_8N	R704HK3	R704I_K3	Match
health_8Distance	R704HK4	R704I_K4	Match
health_8EasytoTravel	R704HK5	R704I_K5	Match
health_9	R704IK2	R704J_K2	Match
health_9N	R704IK3	R704J_K3	Match
health_9Distance	R704IK4	R704J_K4	Match
health_9EasytoTravel	R704IK5	R704J_K5	Match
health_10	R704JK2	R704K_K2	Match
health_10N	R704JK3	R704K_K3	Match
health_11	R704KK2	R704L_K2	Match
health_11N	R704KK3	R704L_K3	Match
health_11Distance	R704KK4	R704L_K4	Match
health_11EasytoTravel	R704KK5	R704L_K5	Match
health_12	R704LK2	R704M_K2	Match
health_12Distance	R704LK4	R704M_K4	Match
health_12EasytoTravel	R704LK5	R704M_K5	Match
posyandu_N	R705B	R705A	Match
posyandu_N2	R705C	R705B	Match
doctorM_N	R707A1	R706A1	Match
doctorF_N	R707A2	R706A2	Match
dentist_N	R707B	R706B	Match
midwife_N	R707C	R706C	Match
otherHealth_N	R707D	R706D	Match
TBA_N	R707E	R708	Match
disease_1	R708AK2	R709A_K2	Match
disease_1N	R708AK3	R709A_K3	Match
disease_1Died	R708AK4	R709A_K4	Match
~	~	~	Match
disease_6	R708HK2	R709H_K2	Match
disease_6N	R708HK3	R709H_K3	Match
disease_6Died	R708HK4	R709H_K4	Match
mulnutrition_N	R709	R710	Match
JAMKESMAS_N	R711	R711A	Match
SKTM_N	R712	R711B	Match

Both datasets contain identical questions on ten common education institutions, seven types of vocational schools, and four types of special education schools (PODES 2014 has five and slightly

different coding scheme). They are broadly matched with each other.

For health facilities, it appears one particular type (PHC) from PODES 2011 is spited into two different types in PODES 2014 (PHC with hospitalization and those without). To retain most information, arithmetic treatment is carried to extract relevant information. All other eleven types are matched perfectly.

Six out of eight outbroken diseases in PODES 2011 are shared in PODES 2014. We retain the common ones and ignore the others.

7. Social and Cultural

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
religion_1	R80101	R80101	Match
~	*~*	*~*	Match
religion_7	R80107	R80107	Match
mostReligion	R802	R802	Recode 2011
worship_1	R803A	R803A	Match
~	*~*	*~*	Match
worship_8	R803H	R803H	Match
disabled_1	R805A	R805A	Match
~	*~*	*~*	Match
disabled_9	R805I	R805I	Match
multiEthnicity	R806	R804A1	Match
commService	R808	R807A	Match

Both datasets share identical questions in regard to seven types of religions, eight types of worship places and nine types of disabilities. The only recode is on the “most popular religion” in the village, where 3 occurrences of “Buddha” in PODES 2011 is recoded to “Other” to match with pattern in PODES 2014 (where there is no “Buddha”). It is also noted that PODES 2014 has not reported major ethnicities.

8. Entertainment

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
cinema	R901A	R902A1	Match
cinemaDistance	R901B	R902A2	Match
pub	R902A	R902B1	Match
pubDistance	R902B	R902B2	Match
sport_1	R903AK2	R903A_K2	Match
sport_1Group	R903AK3	R903A_K3	Match
~	*~*	*~*	Match
sport_7	R903GK2	R903H_K2	Match
sport_7Group	R903GK3	R903H_K3	Match

sport_8Group	R903HK3	R903F_K3	Match
sport_9Group	R903IK3	R903I_K3	Match
sport_10Group	R903JK3	R903J_K3	Match

All common variables matched.

9. Transportation, Communication and Information

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comment*
infrastructure	R1001A	R1001A	Match
roadType	R1001B1	R1001B1	Match
roadThrough	R1001B2	R1001B2	Recode 2014
distance_1	R1004AK2	R1002A_K2	Match
distance_1Transport	R1004AK5	R1002A_K6	Recode 2014
distance_2	R1004BK2	R1002B_K2	Match
distance_2Transport	R1004BK5	R1002B_K6	Recode 2014
distance_3	R1004CK2	R1002D_K2	Match
distance_3Transport	R1004CK5	R1002D_K6	Recode 2014
landPhone	R1005A	R1003A	Match
landPhone_N	R1005B	R1003B	Match
payPhone	R1006	R1004A	Match
BTS	R1007A	R1005A	Match
cellPhone	R1007B	R1005B	Match
telecomm	R1008	R1006	Match
internet	R1009	R1007B	2014 recode to 1 or 2
postOffice	R1010A	R1008A	Match
mobilePO	R1011	R1008B	Match
TV_1	R1012A	R1009B_K2	Match
TV_2	R1012B	R1009C_K2	2011 recode to 1 or 2
TV_3	R1012C	R1009A_K2	2011 recode to 1 or 2
TV_4	R1012D	R1009D_K2	2011 recode to 1 or 2

The road infrastructure variables are mostly shareable between the datasets, however there are some major recodings carried out to keep PODES 2014 consistent with 2011. The different wordings from the original document (in Indonesian) for "***distance_Transport***" are tricky for decision, and therefore requires some subjective judgment. Besides, all other common variables in this section are matched with only small discrepancies.

10. Land Use

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comments*
landChg_1	R1103AK4	R1103A_K4	Match
landChg_1pct	R1103AK5	R1103A_K5	Match
~	~	~	Match

landChg_6	R1103CK4	R1103C_K4	Match
landChg_6pct	R1103CK5	R1103C_K5	Match

All common variables matched.

11. Economy

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comments*
KUD	R1201A	R1213A	Match
nonKUD	R1201B	R1213B	Match
industry_1	R1202A	R1201A	Match
~	*~*	*~*	Match
industry_8	R1202H	R1201H	Match
gpShop	R1203A	R1202A	Match
gpShopDis	R1203B	R1202B	Match
kerosene	R1204A	R1203A	Match
LPG	R1204B	R1203B	Match
prmntMkt	R1205A	R1204A, R1204B	Match
prmntMktDis	R1205B	R1204C	Match
fltMkt	R1206	R1205	Match
miniMkt	R1207	R1206	Match
grocery	R1208	R1207	Match
tavern	R1209	R1208	Match
restaurant	R1210	R1209	Match
hotel	R1211	R1210	Match
lodge	R1212	R1211	Match
cooperative_1	R1213A	R1212A	Match
cooperative_2	R1213B	R1212B	Match
cooperative_3	R1213C	R1212C	Match
cooperative_4	R1213D	R1212D	Match
creditFclty_1	R1214A	R1214A	Match
creditFclty_2	R1214C	R1214C	Match
bank_1	R1215AK2	R1215A_K2, R1215B_K2	Match
bank_1N	R1215AK3	R1215A_K2, R1215B_K2	Match
bank_2	R1215BK2	R1215C_K2	Match
bank_2N	R1215BK3	R1215C_K3	Match

All common variables in this section are matched, with only small change of PODES 2014 (by collapsing two variables into one) to keep consistent with 2011.

12. Security

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comments*
massFght	R1301A	R1301A	Match
fght_1N	R1301B1K2	R1301B1_K2	Match

fght_1Dead	R1301B1K3	R1301B1_K3	Match
fght_1Injured	R1301B1K4	R1301B1_K4	Match
	R1301B1K5	R1301B1_K4	Not used: 2014 messy answer
~	~	~	Match
fght_7N	R1301B7K2	R1301B7_K2	Match
fght_7Dead	R1301B7K3	R1301B7_K3	Match
fght_7Injured	R1301B7K4	R1301B7_K4	Match
	R1302A	R1302A	Not used: 2014 unmatchable answer
	R1302B	R1302B	Not used: 2014 messy answer
criminal_1	R130301K2	R1303A01K3	Match
criminal_1Tndcy	R130301K3	R1303A01K4	Match
~	~	~	Match
criminal_10	R130310K2	R1303A10K3	Match
criminal_10Tndcy	R130310K3	R1303A10K4	Match
suicide	R1304	R1307	Match
streetChild	R1306	R1308A	Recode 2011
sexWorker	R1307	R1309	Match
secActivities_1	R1308A	R1304A	Match
secActivities_2	R1308B	R1304B	Match
secActivities_3	R1308C	R1304C	Match
secActivities_4	R1308D	R1304D	Match
police	R1309BK2	R1306A	Match
policeDist	R1309BK3	R1306B1	Match
policeEasytoTravel	R1309BK4	R1306B2	Match
guards	R1310	R1305	Match

All common variables are generally matched in this section, except for minor recodings. It is noted that some variables, though matched according to survey document, presents vague or messy answers in the datasets. They are therefore not used for global schema.

13. Autonomy

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comments*
income_1Form	R1401AK2	R1501A_K2	Match
income_1	R1401AK3	R1501A_K3	Match
~	~	~	Match
income_7Form	R1401B6K2	R1501C6_K2	Match
income_7	R1401B6K3	R1501C6_K3	Match
dev_1	R1402A1K2	R1401A1_K2	Match
dev_1Source	R1402A1K3	R1401A1_K3	Match
~	~	~	Match

dev_4	R1402A4K2	R1401A4_K2	Match
dev_4Source	R1402A4K3	R1401A4_K3	Match
fund_1	R1402B1K2	R1401B1_K2	Match
fund_1Source	R1402B1K3	R1401B1_K3	Match
~	~	~	Match
fund_3	R1402B3K2	R1401B3_K2	Match
fund_3Source	R1402B3K3	R1401B3_K3	Match
progm_1	R1402C1K2	R1401C1_K2	Match
progm_1Source	R1402C1K3	R1401C1_K3	Match
~	~	~	Match
progm_3	R1402C3K2	R1401C3_K2	Match
progm_3Source	R1402C3K3	R1401C3_K3	Match

All common variables matched.

14. Village Administration

Global Schema	*Correspondent 2011*	*Correspondent 2014*	*Comments*
head	R1501AK2	R1601A_K2	Match
headAge	R1501AK3	R1601A_K3	Match
headSex	R1501AK4	R1601A_K4	Match
headEdu	R1501AK5	R1601A_K5	Recode 2014
sec	R1501BK2	R1601B_K2	Match
secAge	R1501BK3	R1601B_K3	Match
secSex	R1501BK4	R1601B_K4	Match
secEdu	R1501BK5	R1601B_K5	Recode 2014

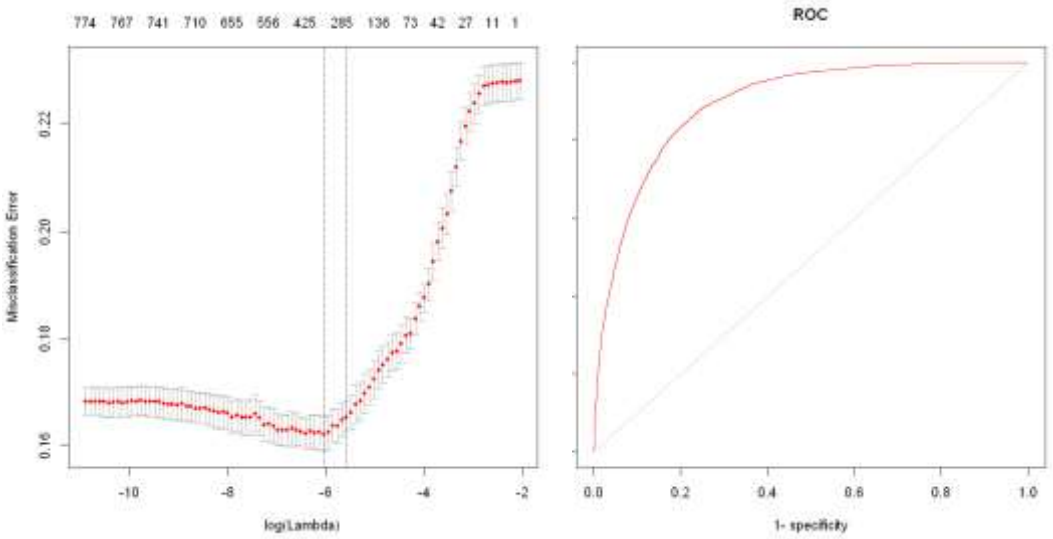
All common variables matched except for minor recoding.

Part V. Modelling and Prediction

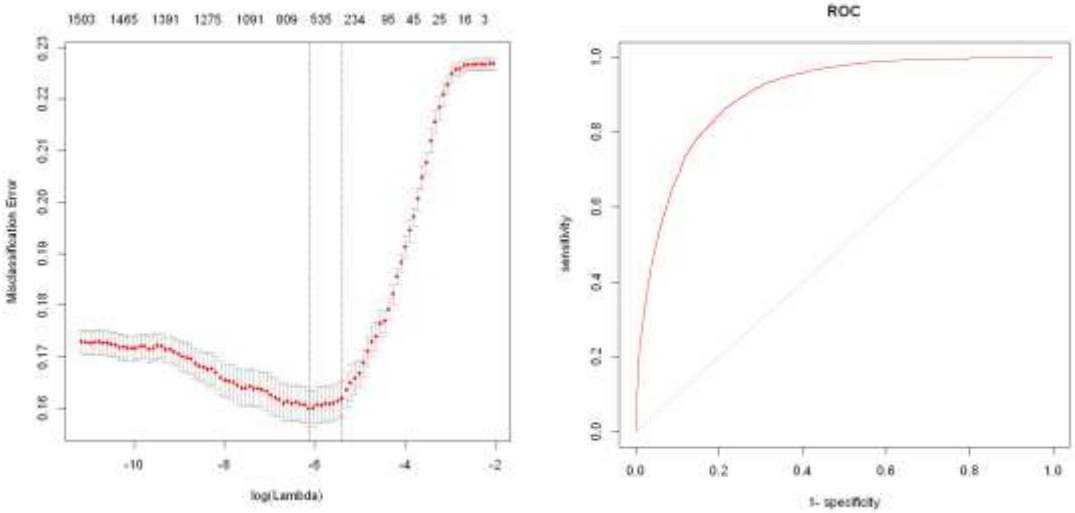
[Modelling]

For a simple and probability-based model, we use Logistic regression with L1 penalty. This is primarily because of its simplicity and the ability to choose features on its own. Given the lack of domain knowledge to build the model from the bottom up, this characteristic is extremely helpful. To control for overfitting, we use cross-validated misclassification error to find the best parameter.

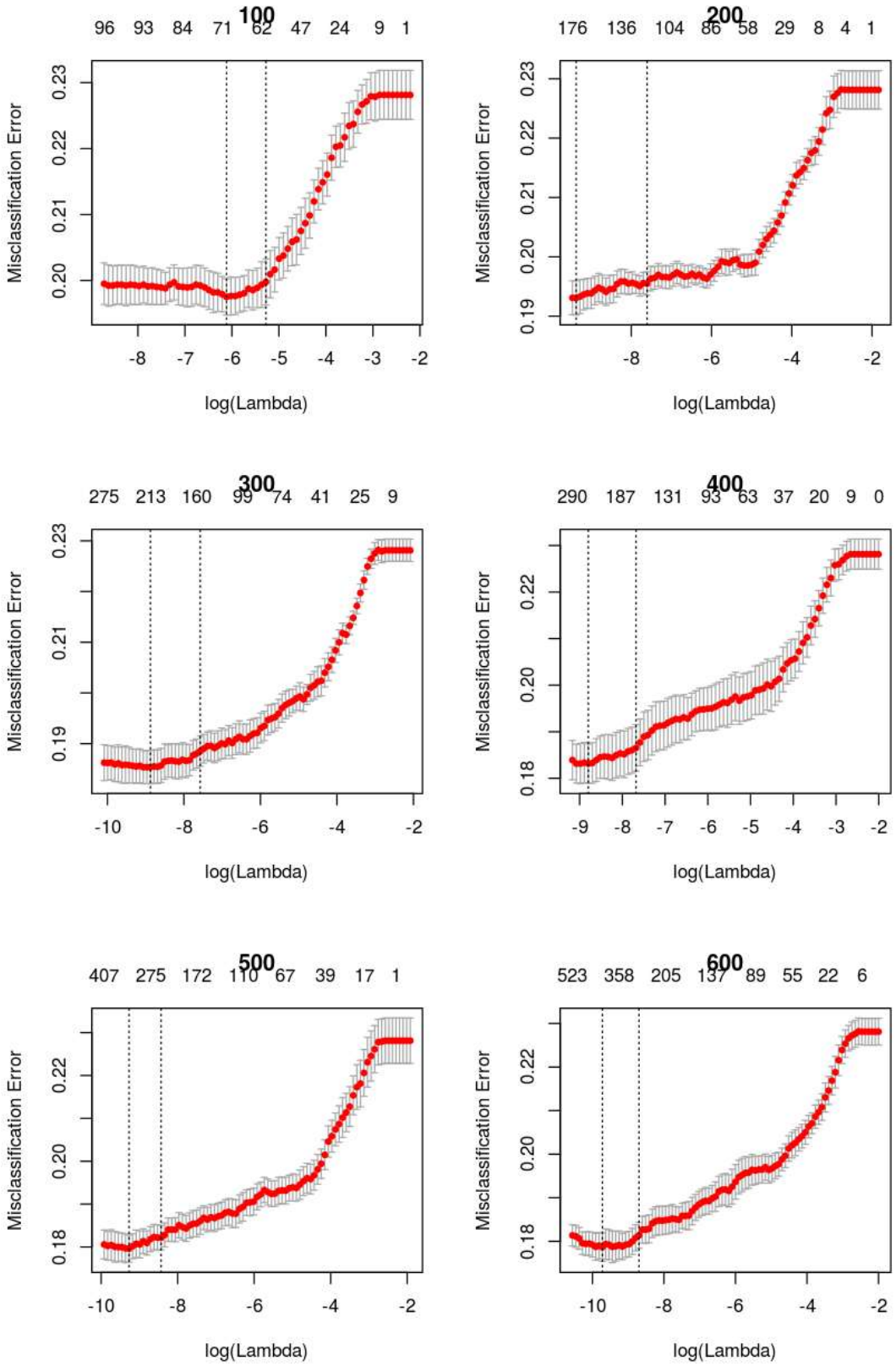
Using a global schema on the training set, the cross-validated misclassification error is 16.2%. Considering there are 3307 "***connect***" villages in the total of 14,496, a blind guess would result in 22.8% error rate. Therefore, the model improves the accuracy by 6.4%. With a one-standard-deviation lambda, 286 variables are reported to have non-zero coefficient. The AUC score is 0.896.



As a comparison, we train the model on the original PODES dataset, which contains 540 useful variables. The misclassification rate slightly decreases to 16.0% with a slight higher AUC = 0.902. Using one-standard-deviation lambda, 349 variables are reported non-zero, which reflects a more complex model.



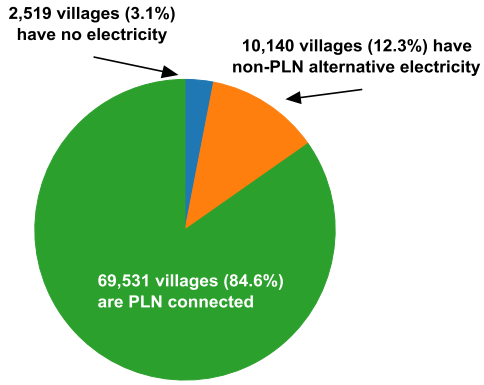
A neural network auto-encoder for self-feature generation is also tested to see if self-generating features (through an auto-encoder) can serve as better predictors than original ones in logistic regression. To save computing power, the network only contains one layer with 50 times of iterations. The following graphs show the cross-validated error rate with different number of neurons in each test. The maximum number is 600, which is less than the original model matrix generated from the dataset.

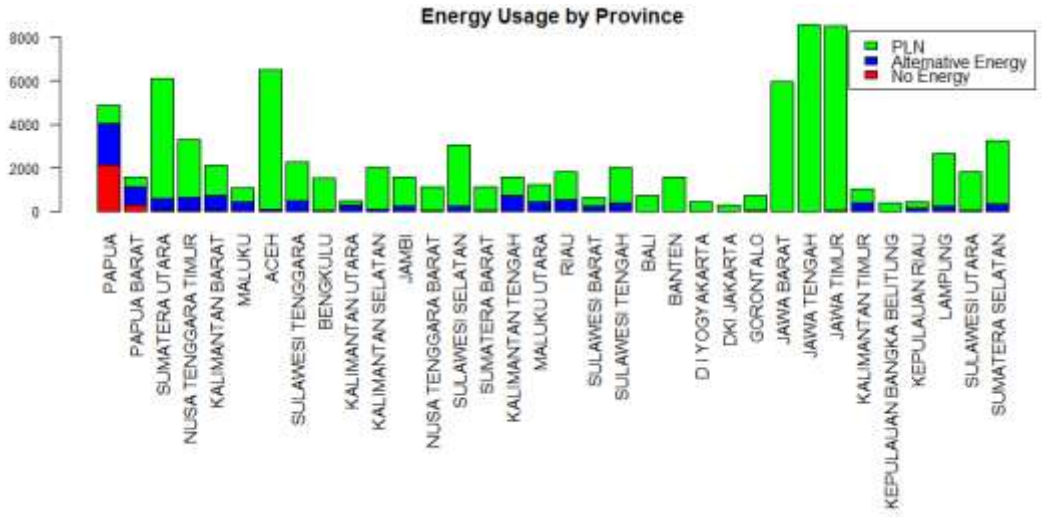


It is note that though following a downward trend, the misclassification rate from the auto-generated features perform no better than the original predictors. Therefore we are comfortable with the original model.

[Prediction]

An overlook of 2014 PODES survey shows around 85% of all national villages are already connected to PLN network. Among 12,659 villages that are not PLN-connected, 2,519 (3.1%) have no electricity at all. We also see there is a wide unbalance between provinces in the number of disconnected villages, with New Guinea having the largest number of villages with no electricity of any kind.





Using the model trained with global schema, we calculate the predicted probability of village connect to PLN on the PODES 2014 dataset and plot them by main islands of Indonesia.

Part VI. Discussion and Next Steps

The maps below show the predicted probability of connecting to PLN by 2017 (a prediction made based on training data in 2011-2014). The darker blue area shows the region that are already PLN-connected. A red-blue color threshold indicates the value of probability that are suitable for each island. All provinces except Java (color threshold = 0.5) and New Guinea (color threshold = 0.05) has the value of 0.1, indicating the red area are less likely to be connected (probability less than 10%).

A preliminary reading of the maps indicates a wide provincial disparity of power connections. In

Java, only small pockets of area are shown in red, showing a near 50% predicted probability of future connection. On the other hand, there are vast areas in Kalimantan and New Guinea that have low probability of future connection. In particular, Papua has the largest coverage of areas that are extremely unlikely to be connected (less than 5%).

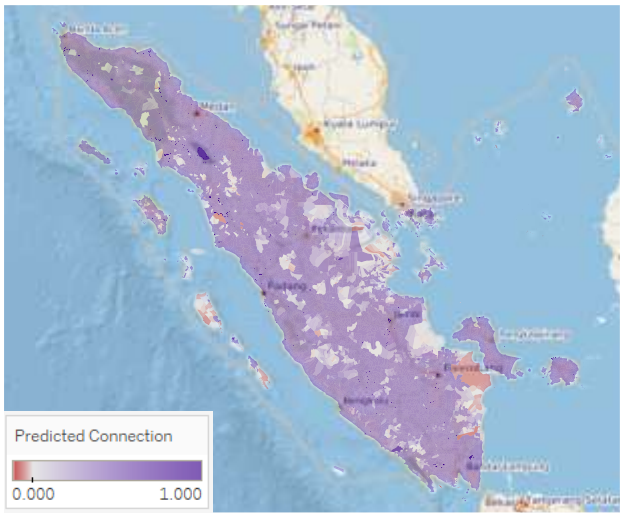
It's also interesting to note that most villages with high predicted probability (light blue in color) are closer to the villages that are already PLN-established, showing geo-location, such as province, district and sub-district play an important role in the predictive model.

We may also observe that current PLN network seems to expand along major highways. Areas that are further away from these roads tend to be less likely to connect in the future.

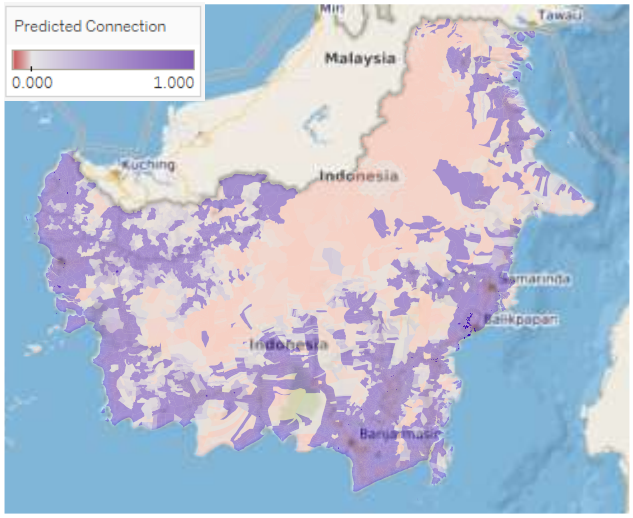
Finally, smaller and separated islands that are away from the main island seem less likely to be connected in the future. It is possible due to their remote location from established network and poorer infrastructure.

JAVA(50%)

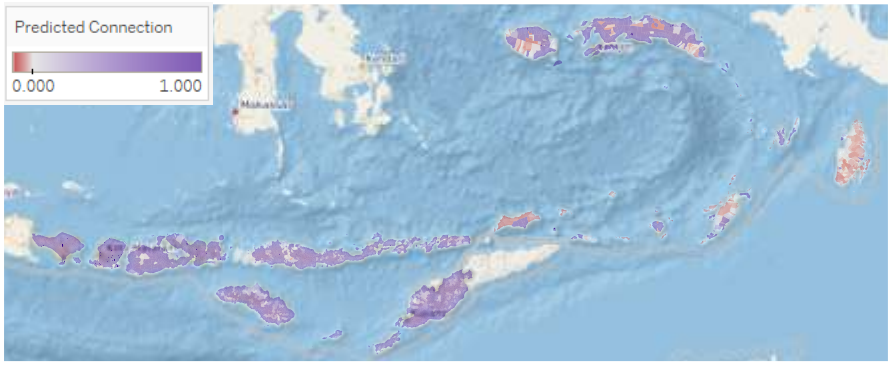
SUMATERA



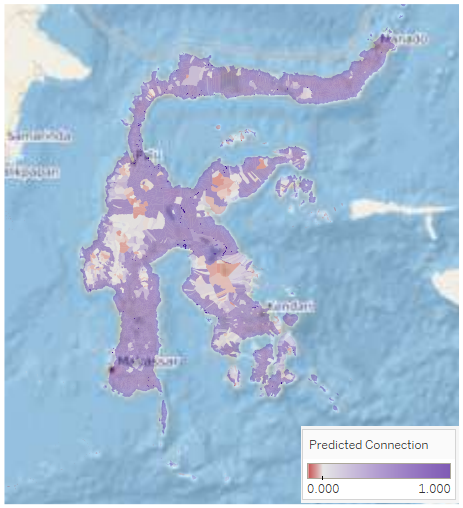
KALIMANTAN



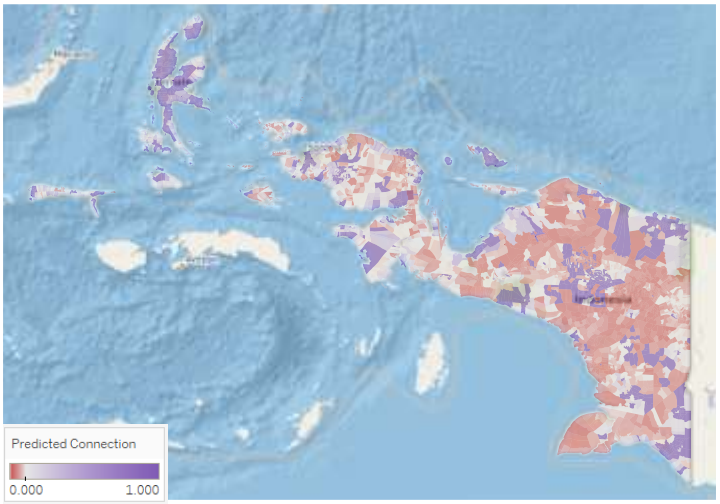
LESSER SUNDA ISLANDS



SULAWESI



NEW GUINEA (5%)



[Next Steps]

It is noteworthy that a few more steps could be taken to refine the model and prediction.

Firstly, we realize that the prediction for New Guinea could be more optimistic than reality. This is primarily because of the treatment of unique sub-districts in PODES 2014 that are not available for PODES 2011. If we can recode them to known sub-district IDs (like North Kalimantan), then we should have improved prediction for the province.

Secondly, due to time restraints and lack of domain knowledge, we have not conducted data imputation in great detail. This results in many sparse variables with large amount of unique level coded as “NA” or 0 (even they are truly a unique level). While this has limited effect on the predicted probability, they do report large coefficients and create difficulty for model interpretation. In order to gain more precise and interpretable insights from the coefficients, it would be a good idea to consider collapsing them into single variables.

Thirdly, even though neural network with auto-encoder did not provide stronger training accuracy than the original model, in the interest of further improve prediction, it is still worthwhile to run

further tests particularly with larger number of iterations. Due to computing limitation, only 50 iterations were used during my tests (maximum would be the number of training observations). It is important to allow for more parameter updates in order to verify if the technique could provide any improvements as a benchmark.

Finally, the global schema could be expanded to meet future survey templates (i.e., 2017) by not deleting unused variables in the training phase. We had deleted these on purpose for faster calculation and a cleaner schema.