Australian National University

Micro-econometrics of Development: A Review

Budy P. Resosudarmo

Indonesia Project
Arndt-Corden Department of Economics
Crawford School of Public Policy
ANU College of Asia & the Pacific

Basic (previous) Model

- Main econometric model

$$y_{i,t} = \alpha_0 + X_{i,t} \times \alpha_1 + e_{i,t}$$

- Derived from household or firm behaviour model (*i* = individual/household/firm)
- *X* = a vector of assumed "exogenous" variables determining *y*
- Conducted in mostly a cross section analysis; some in a panel data analysis → solving time-invariant issue
- Difficulties in ensuring the exogeneity of X
- Endogeneity issues
- Association vs Causality? Mostly only correlation

Why causality is important?

- For policy-makers, what is important not so much whether a policy associated with a certain outcome
- Policy-makers want to know if a policy is implemented how much it will affect a certain outcome

$$\nabla y_{i,t} = \alpha . \nabla p_{i,t}$$

- To get this answer, you need to have a causality relation between the policy and the outcome
- An association relation will not produce this answer
- Benarjee, Duflo, Karlan and the J-PAL group strongly argued on this direction

Basic (current) Model

- Main econometric model

$$y_{i,t} = a_0 + X_{i,t} \times a_1 + a_2 \times p_{i,t} + e_{i,t}$$

- Focus on *p* = a certain policy variable affecting *i*
- Resolve the endogeneity issues:
  - Omitted confounding variable: other variables not in the model
  - Measurement error: some variables are typically not reliable
  - Simultaneity
- Argue for causality relation
- This approach is argued to be more useful for policy-makers

Endogeneity I: Omitted Confounding Variable

- Suppose the true model is:

$$y_{i,t} = a_0 + X_{i,t} \times a_1 + a_2 \times p_{i,t} + a_3 \times z_{i,t} + e_{i,t}$$

- We omit $z_{i,t}$, then our model:

$$y_{i,t} = a_0 + X_{i,t} \times a_1 + a_2 \times p_{i,t} + e_{i,t} \text{ where } e_{i,t} = a_2 \times z_{i,t} + u_{r,t}$$

- If *p* <u>has some correlation </u>with *z* and *z* separately affects *y*, then our estimate of

$$y_{i,t} = \alpha_0 + X_{i,t} \cdot \alpha_1 + \tilde{\alpha}_2 \cdot p_{i,t} + e'_{i,t}$$

- produces $\tilde{\alpha}_2 \neq \alpha_2$

Endogeneity II: Measurement Error

- Suppose we can only observe *p**, where:

$$p^*_{i,t} = p_{i,t} + v_{i,t} \text{ where } \mathrm{v}_{i,\mathrm{t}} \text{ is "noise"}$$

- then our model: $y_{i,t} = \alpha_0 + X_{i,t} \cdot \alpha_1 + \alpha_2 \cdot p^*_{i,t} + e_{i,t}$

$$\text{having } e_{i,t} = -\alpha_2 \cdot v_{i,t} + u_{i,t}$$

- If *v* is not random, then our estimate of

$$y_{i,t} = \alpha_0 + X_{i,t} \cdot \alpha_1 + \tilde{\alpha}_2 \cdot p^*_{i,t} + e'_{i,t}$$

- produces $\tilde{\alpha}_2 \neq \alpha_2$

Endogeneity III: Simultaneity

- Suppose that for some reason we believe that:

$$p_{i,t} = g_0 + g_1 \times y_{i,t} + u_{i,t}$$

- In this case our p and e will be correlated and so our estimate of $y_{i,t} = \alpha_0 + X_{i,t} \cdot \alpha_1 + \tilde{\alpha}_2 \cdot p_{i,t} + e'_{i,t}$

- produces $\tilde{\alpha}_2 \neq \alpha_2$

Empirical Strategy

- **<u>Identification Strategy</u>**:

  The part that you argue or ensure that your p is a random and exogenous variable

  - The way that p is assigned to each individual is not co-determined with the same variable that is not controlled
  - Ensuring p is not correlated to the error term

- **<u>Main Estimations</u>**:

  Explaining your main basic model and results of your estimation

- **<u>Robustness tests</u>**:

  The part that you need to test that your result are robustly not contained any endogeneity issue

- **<u>Extended analysis</u>**:
  - Heterogeneity analysis
  - Channels

Identification Strategy I

By Experiments → Randomized Control Trials

N_T — Treatment group = receiving policy p

N_C — Control group = not receiving policy p

$$ATE = \hat{y}_T - \hat{y}_C$$

- Issues:
  - External validity
  - Spillover and/or crossover effects
  - Attrition bias
  - You know the impact, but not the optimal policy
- Important to have a proper design of samples
  - List of the population being researched
  - Take a random sample at the unit being analyzed from the list

Identification Strategy II

Natural Experiments

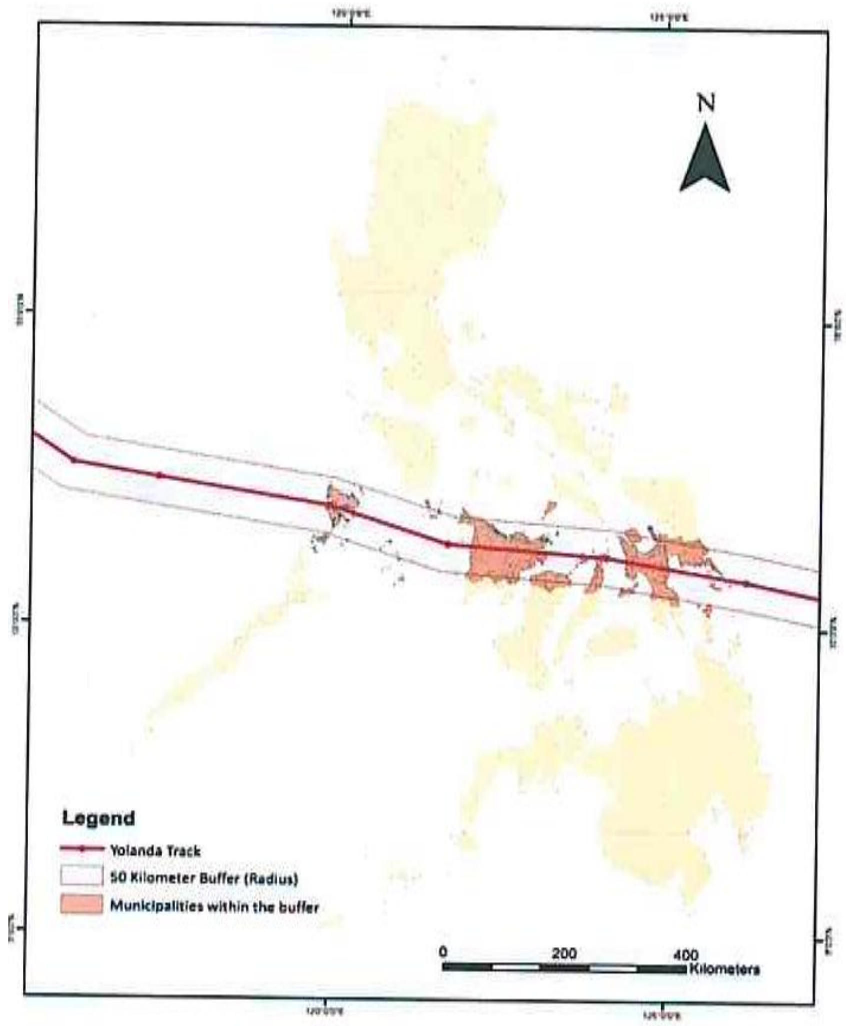
- Also known as one of the quasi experiments
- Not an experiment, but we set it the events such as they are a random event
- Which is: being treated or not is determined by nature

- The most common cases are to analyze the impact of natural hazard event on human welfare

- Issues: the magnitude of natural hazard event might be random but the location and timing often are predictable and so the location of people affected is not random
- Even if the event of natural hazards is purely random, what is the policy?

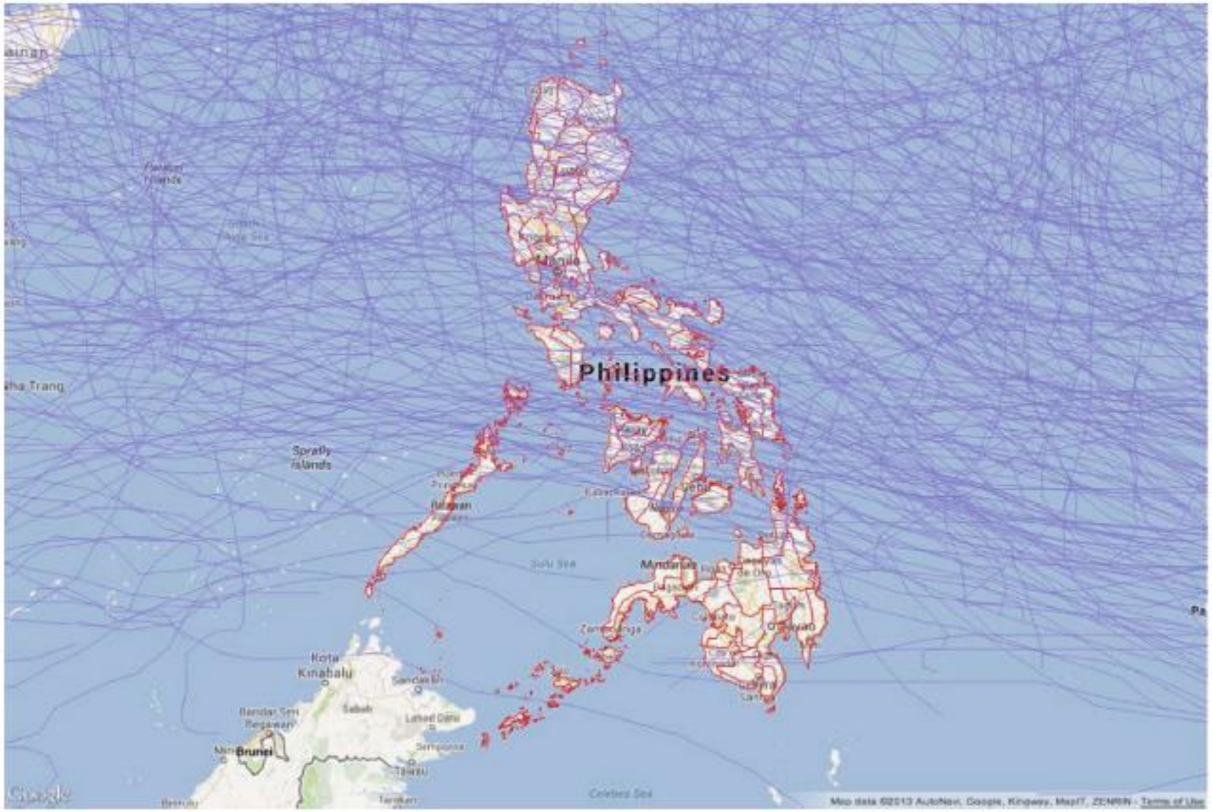
Example: Typhoon Haiyan

- Typhoon entered the Philippine Area of Responsibility (PAR), November 6, 2013
- Typhoon intensified moving West-Northwest towards Eastern Visayas, November 7, 2013
- Made 1st landfall over Eastern Samar (Region 8), 2nd Landfall over Leyte (Region 8, and then towards Cebu (Region 7), November 8, 2013
- Exited PAR November 9, 2013
- One of the strongest and disastrous tropical cyclones, maximum sustained winds 195kph near center and gustiness of 230kph upon entering PAR, moving WNW at 30kph



Australian National University

Identification Strategy

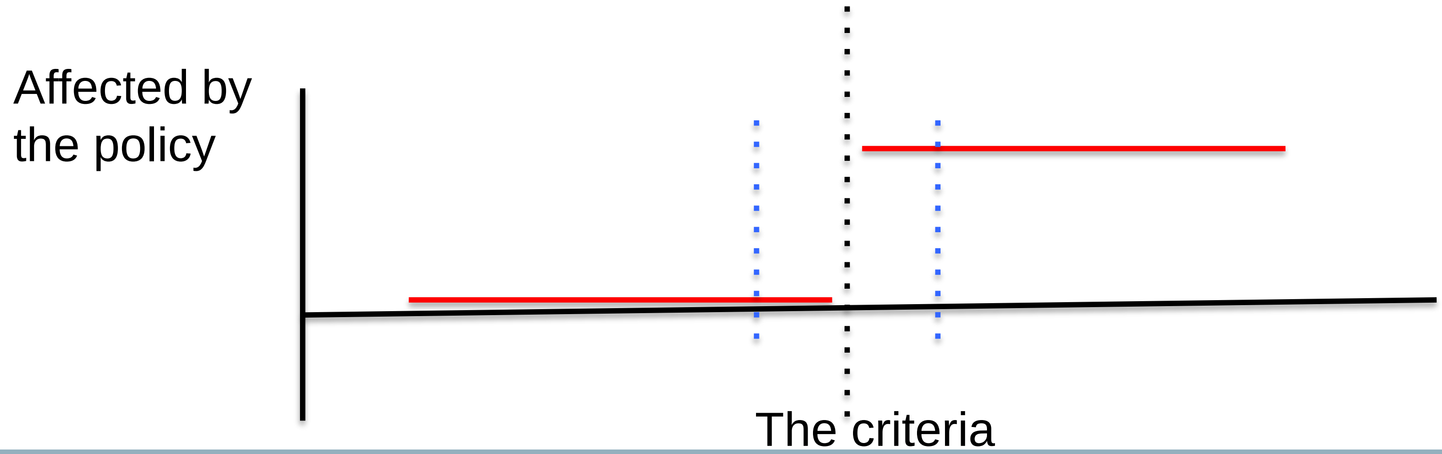


- Tracks of tropical storms (1990-2010)
- Licuanan, Mahmoud, and Steinmayr (2014). Data comes from the National Climatic Data Centre (NCDC): http://www.ncdc.noaa.govlibtracslindex.php?name=ibtracs-data. Base map comes from Google Maps.

Identification Strategy III

Discontinue Event

- Another type of quasi experiments
- The implementation of the policy is not random, it is actually targeted to a very strict rule (discontinuity event)
- The random event comes from the fact that those just not able to fulfill the rule and those just fulfill the rule are the same entities. It is just a random event that one can fulfill and one cannot
- Issues: Need to have discontinuity on the outcome; external validity; exclusivity, defining the cut-off

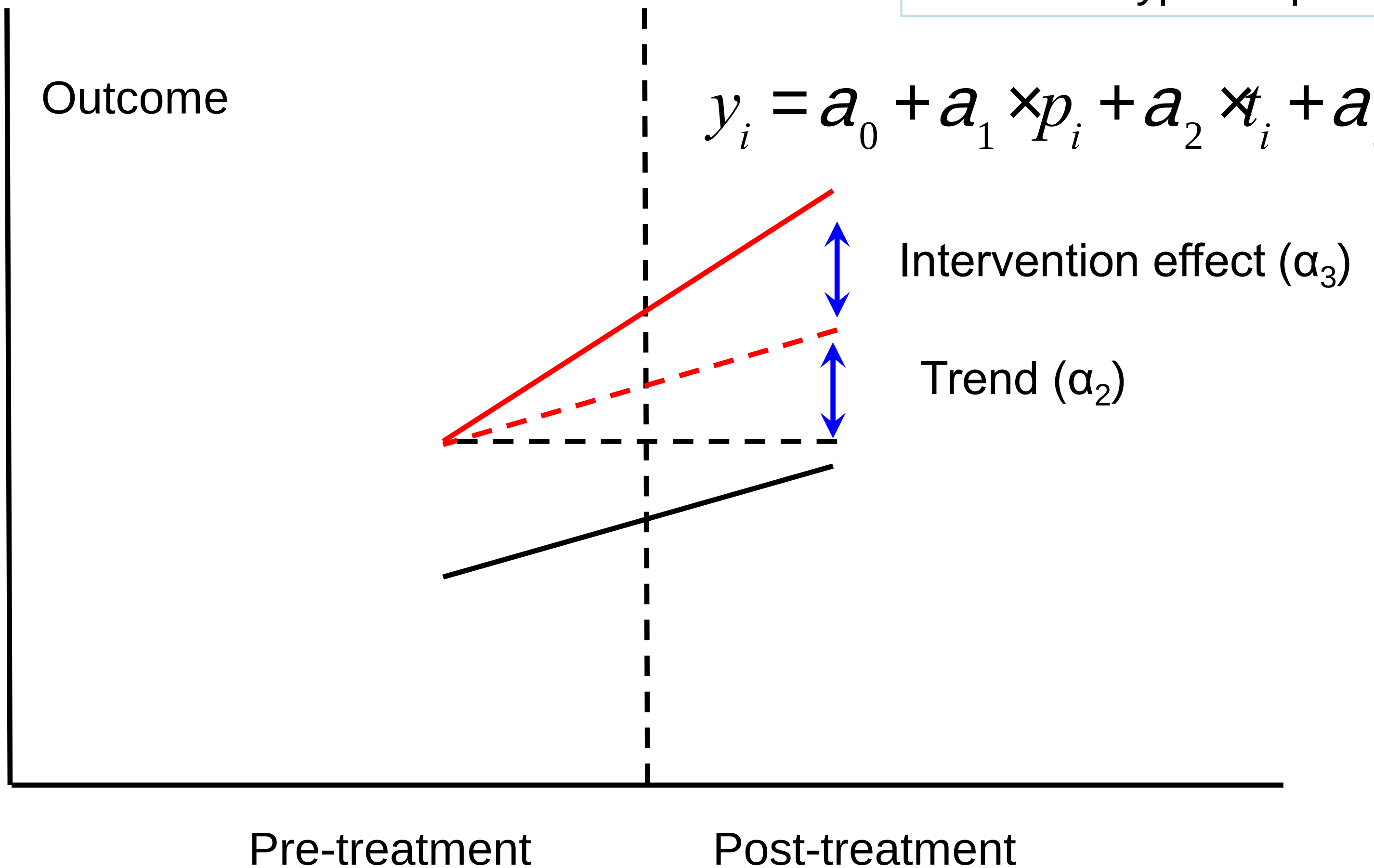


Identification Strategy IV

Difference in Differences

* Another type of quasi-experiment

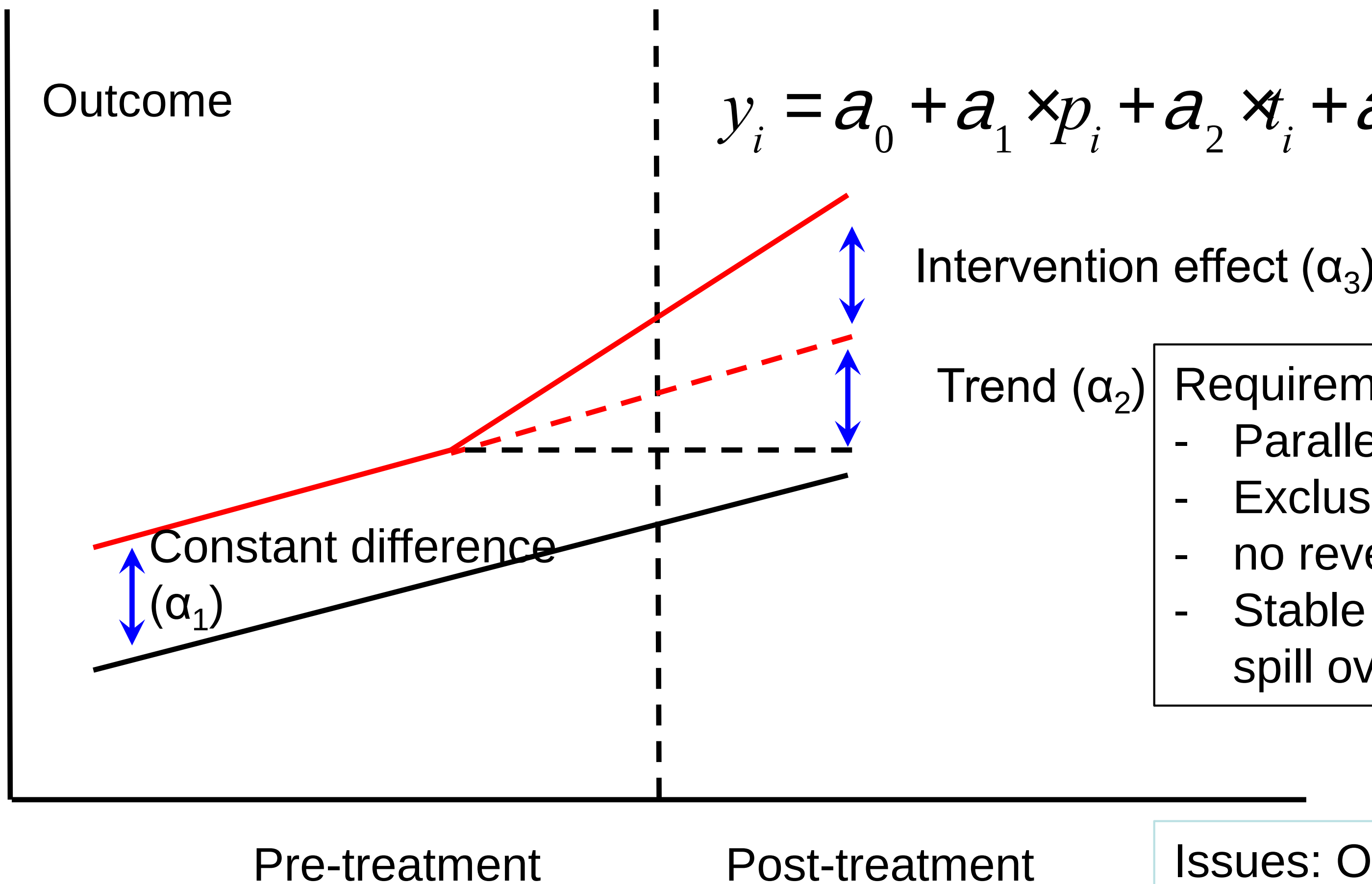
$$y_i = a_0 + a_1 \times p_i + a_2 \times t_i + a_3 \times t_i \times p_i + e_i$$



Identification Strategy IV

Difference in Differences

$$y_i = a_0 + a_1 \times p_i + a_2 \times t_i + a_3 \times t_i \times p_i + e_i$$



Requirements:
- Parallel trend
- Exclusive treatment
- no reverse causality
- Stable group and no spill over

Issues: Omitting variables; reverse

Identification Strategy VI

Panel Fixed Effect Technique

* Another type of quasi-experiment

- Traditional fixed effect technique

$$y_{i,t} = \alpha_0 + X_{i,t}.\alpha_1 + \alpha_2.p_{i,t} + \delta_i + e_{i,t}$$

  - $\boldsymbol{\delta}_I$ = unobserved time-invariant individual effect
  - Control time-invariant variables
  - Issues: time-variant omitting variables; reverse causality
- Modified fixed effect technique

$$\nabla y_{i,t} = \alpha_0 + \nabla X_{i,t}.\alpha_1 + \alpha_2.\nabla p_{i,t} + e_{i,t}$$

  - Could control for initial condition and higher unit of fixed affect; less efficient but could have tighter control

Identification Strategy V

Instrumental Variable Technique

- First-stage

$$p_{i,t} = b_0 + X_{i,t} \times b_1 + b_2 \times z_{i,t} + v_{i,t}$$

- Second-stage

$$y_{i,t} = a_0 + X_{i,t} \times a_1 + a_2 \times \hat{p}_{i,t} + e_{i,t}$$

- Requirements:
  - Cov (p,z) ≠ 0 (first stage exists)
  - Cov (z,e) = 0 (exclusion restriction: z is uncorrelated with any other determinants of the dependent variable)

Robustness Tests

- No perfect test, but you have to try
- For randomized control trials:
  - Fail to control confounding variables
  - Fail to have an external validity
- For quasi-experiments:
  - Fail to control confounding variables omitting variable issue
  - Fail to control reverse causality
- For instrumental variable technique
  - Fail to get a significant exogenous and random instrument
  - Fail to fulfill the property of exclusion restriction

Robustness Test I

Omitting confounding variable issues

- Placebo/Falsification test
  - Regress the equation among control group and group(s) that we know do not receive any treatment
  - Regress the equation among control and treatment group before receiving any treatment
- For example, the case of DiD:

$$y_i = a_0 + a_1 \times p_i + a_2 \times t_i + a_3 \times t_i \times p_i + e_i$$

but for t= -1 and 0

Or between 2 untreated groups

Australian National University

The Impact of Hunger in Java

$$Y_{i,r,t} = \alpha + \beta_1 . c_{i,t} + \beta_2 . h_{i,r} + \beta_3 . (c_{i,t} . h_{i,r}) + \theta . X_{i,r,t} + \varepsilon_{i,r,t}$$

Net effect

Outcomes:

- Years of schooling

Cohort groups:

- Born 1962-1969
- Born 1970-1974 (control)

- Hunger regions
  - Java
- Non hunger region (control):
  - Sumatera

X = **demography**; **parents characteristics**; **household characteristics; region characteristics**.

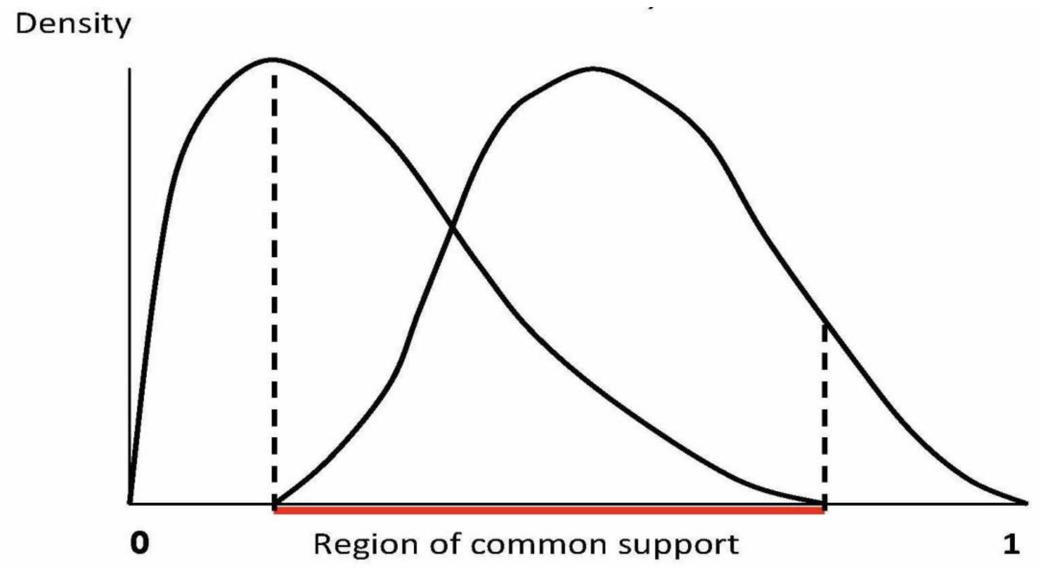
Parallel trend assumption

- In the absence of treatment (hunger), people in the affected and non-affected regions should have similar trend in outcomes
- Two assessments:
  - Using same regions (J, B, NT vs Sumatera) but fake cohort (non hunger cohorts): born in <u>1975-1980</u> vs 1970-1974
  - Using same cohorts (1956-1961, 1962-1969, 1970-1974) but fake treatment birthplace: (<u>Kalimantan</u> vs Sumatera)
- Should expect:
  - no significant difference in trend between hunger affected region and control in the non-hunger periods (Cohorts 1970-1974 & 1975-1980)
  - no significant difference in trend between non-affected regions (Sumatera & Kalimantan)

Robustness Test III

Omitting confounding variable issues

- Apply matching procedure before the regression --> another type of quasi-experiment technique
- Pick controlled observations that “match” treated observations and vice versa



Some methods employ weighting procedures

Robustness Test III

Omitting variable issues

- Apply matching procedure before the regression --> another type of quasi-experiment technique
- Variables to be matched usually are those not correlated with the outcome (but could be different strategy)
- Several matching methods: Propensity Score Matching, Exact Matching and Coarsened Exact Matching
- Issues:
  - need more samples; larger than RCT
  - Could create externality problem

The Case of Typhoon Haiyan

Estimates of remittance participation using OLS and CEM weight-adjusted OLS

	OLS				CEM^^			
Variable	(1)	(2)	(3)	(4)	(9)	(10)	(11)	(12)
Treatment								
DID	**0.0384 *****	**0.0333 *****	**0.0331 *****	**0.0361 *****	**0.0352 ****	**0.0327 ****	**0.0328 ****	**0.0362 *****
Affected	-0.0331 ***	-0.0021	0.0022	-0.1674 ***	-0.0100	0.0071	0.0058	-0.1582 ***
Year	-0.0015	-0.0067	-0.0051	-0.0053	0.0352	-0.0033	-0.0044	-0.0050
Controls^^^								
Wealth	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Head Education	No	No	Yes	Yes	No	No	Yes	Yes
& F. Size								
Region	No	No	No	Yes	No	No	No	Yes
R-squared	0.0007	0.0563	0.0669	0.0824	0.002	0.0166	0.0192	0.0332
Observations	21,333	21,333	21,333	21,333	20,957	20,957	20,957	20,957

Notes: **, *** indicates significance at 5% and 1% levels, respectively.
^^Matching based on family size, household head educational attainment and age, and owning house and lot. Before matching Multivariate L1 distance = 0.4811; after matching Multivariate L1 distance: 0.4068; Number of strata = 428; Number of matched strata = 324. Unmatched in control = 354, unmatched in treatment = 22.
^^^Wealth controls include total income and a vector of dummies pertaining to roof materials and wall materials (strong, light); head education controls include vector of dummies pertaining to household head's highest grade completed (attainment according to Table 4); family size; region controls pertain to region fixed effects.

Robustness Test IV

Omitting confounding variable issues

- Test by adding variables that presumably not correlated with *p* and see whether α_2 stable

$$y_{i,t} = a_0 + X_{i,t} \times a_1 + a_2 \times p_{i,t} + a_3 \times z_{i,t} + e_{i,t}$$

- $Z_{i,t}$ = a vector of variables expected not confounding variables

- Or apply the Oster test (see Stata manual)

Australian National University

Robustness Test V

Reverse causality issues

- Test by adding lag dependent variable that presumably not correlated with *p* and see whether $\boldsymbol{\alpha}_2$ stable

$$y_{i,t} = \alpha_0 + X_{i,t}.\alpha_1 + \alpha_2.p_{i,t} + \alpha_3.y_{i,t-1} + e_{i,t}$$

- Again not a perfect test of reverse causality bias

Robustness Test VI

For the case of Instrumental Variable Technique: Exclusivity restriction

- Test by adding the instrumental variables that presumably not correlated with *p* and see whether $\boldsymbol{\alpha}_2$ stable and $\boldsymbol{\alpha}_3$ is not significant

$$y_{i,t} = a_0 + X_{i,t} \times a_1 + a_2 \times p_{i,t} + a_3 \times z_{i,t} + e_{i,t}$$

- $Z_{i,t}$ = a vector of instrumental variables
- Or regress $Z_{i,t}$ on the estimated error terms or the main equation model

Extended Analysis

- Heterogeneity analysis:
  - to see whether or not the impact of the policy uniform within groups in the treated group
- Channels:
  - to understand through which the policy affects the outcome
  - Step 1:

$$m_{i,t} = \alpha_0 + X_{i,t}.\alpha_1 + \alpha_2.p_{i,t} + e_{i,t}$$

  - $Z_{i,t}$ = a vector of channel variables
  - Step 2:

$$y_{i,t} = \alpha_0 + X_{i,t}.\alpha_1 + \alpha_2.m_{i,t} + e_{i,t}$$

Australian National University

Using Big Data for Social Analysis in Indonesia: Opportunities and Challenges

Budy P. Resosudarmo

Indonesia Project
Arndt-Corden Department of Economics
Crawford School of Public Policy
ANU College of Asia & the Pacific

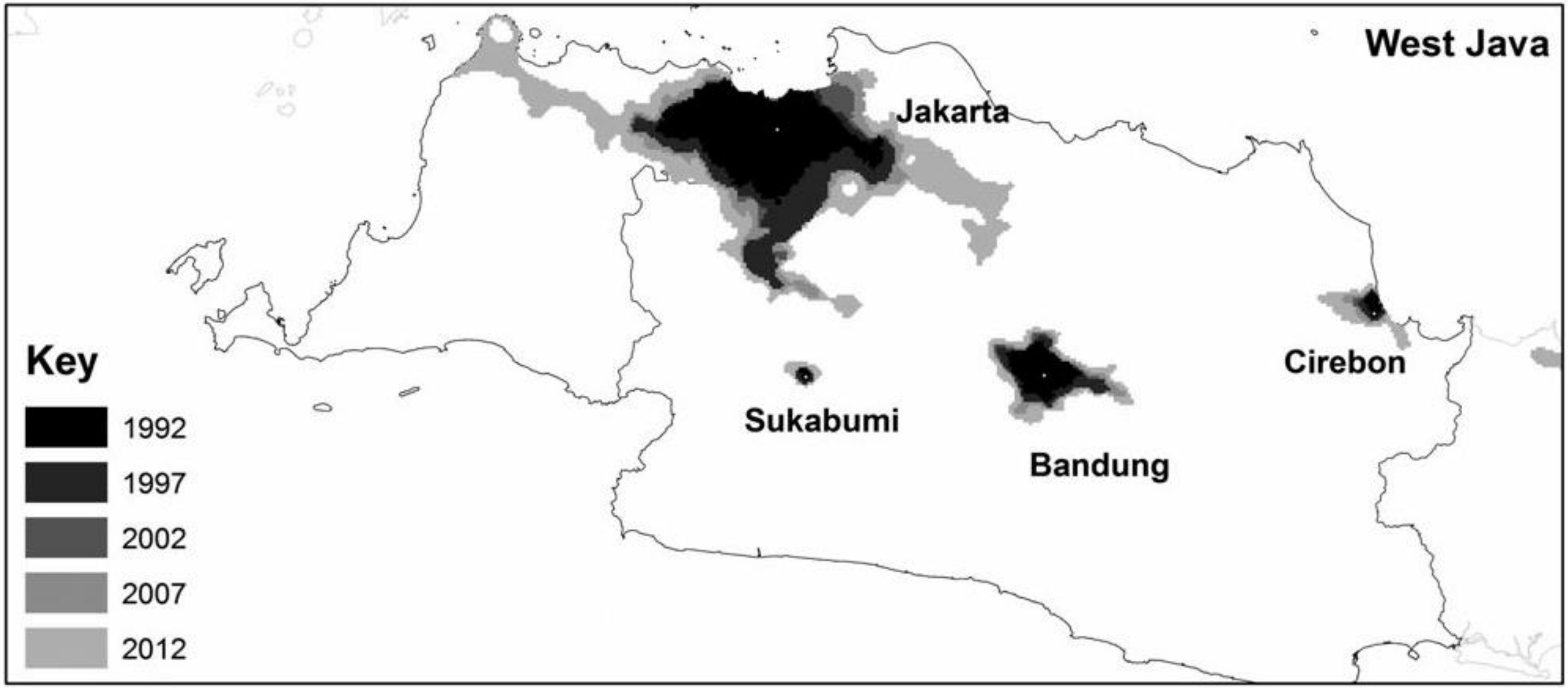
Big Data

- Big data: an umbrella term for large digital data continuous collected from the global population through utilization of digital technologies.
  - It contains information related to human behaviour and interactions.
- 3 v's: volume, velocity, variety (+ value)
  - social media contents, online searches (private citizen)
  - digital opinion poll, online/mobile apps (private citizen and businesses, locations)
  - tracking data from mobile devices (usage patterns, locations)
  - sensor network sources: satellite imaging, weather/climate sensors (bio-physical condition, locations)
  - commercial transactions (between individuals and businesses)
  - electronic administrative records, bank records, insurance records, toll road utilization (private sector, government)
- New source of (massive) information; including for social scientists

Traditional and Big Data

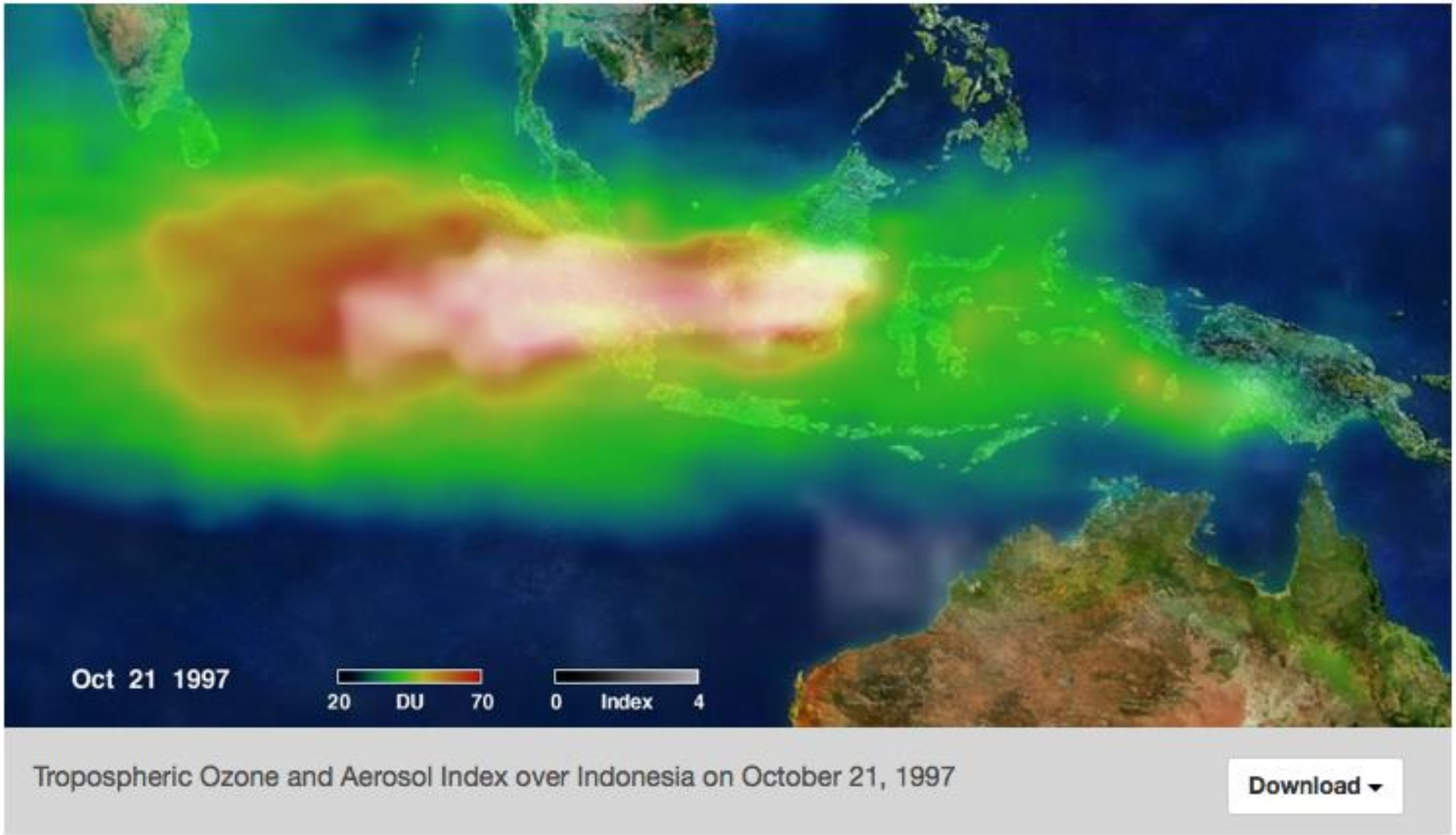
- Traditional Data
  - Contained individual and household information
  - Hypothesis driven
  - Multi topics
  - Actively collected
  - Selected random sample
  - Possible to have a longitudinal information
  - Sample size: small to large
- Big Data
  - Contained mostly individual and, in few cases, household information
  - Opportunity driven
  - Mostly a particular topic
  - Passively collected
  - Tend to have a bias sample
  - Mostly cross-section information; few cases to have a longitudinal information
  - Sample size: large to very large

Night light data for detecting urban sprawl



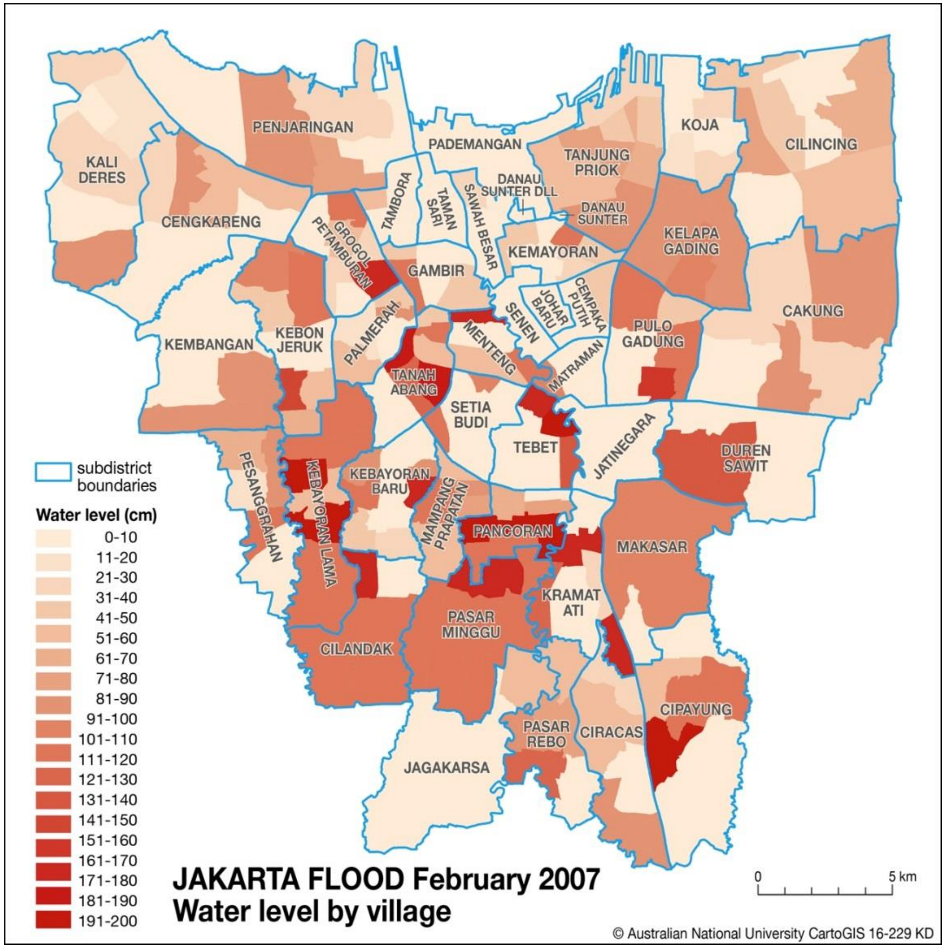
"Urban land expansion in Indonesia 1992-2012: Evidence from satellite-detected luminosity" by Olivia, S., G. Boe-Gibson, G. Stitchbury, L. Brabyn and J. Gibson

Aerosol index to measure the impact of tariff reforms



“Regional tariff reform exposure and air quality in Indonesia” by Yessi Vadilla and Budy P. Resosudarmo

Flood and the value of house



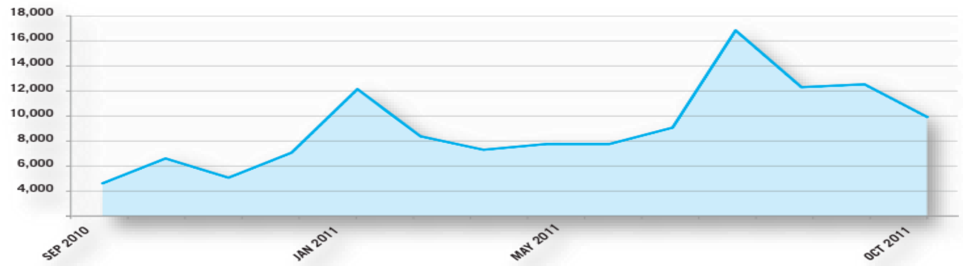
“The cost of floods in developing countries’ megacities: A hedonic price analysis of the Jakarta housing market, Indonesia” by José Cobián Álvarez and Budy P. Resosudarmo

Tweets predict food basket inflation

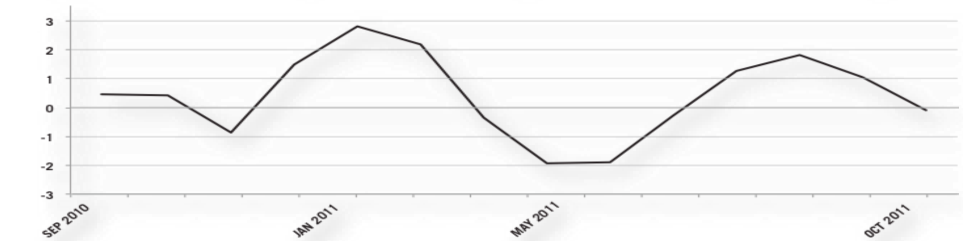
(rice, chilies, fish, sugar, corn, cooking oil)

Tweets about the price of rice

(per month)

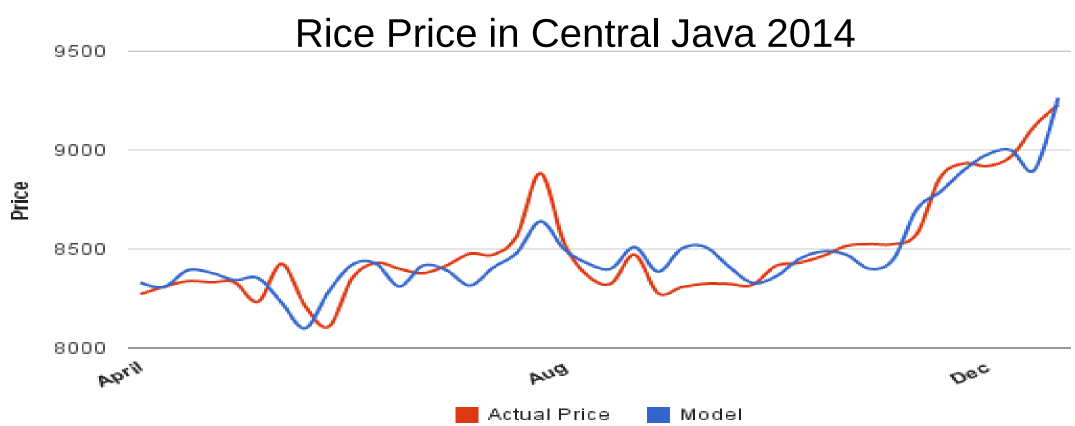


Official Food Price Inflation
(monthly from 25 cities)



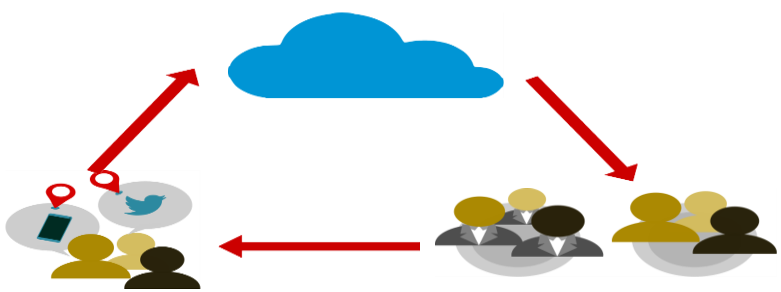
Predicting food prices
Provincial Price = Past National Price
+ Recent Google Trends in province

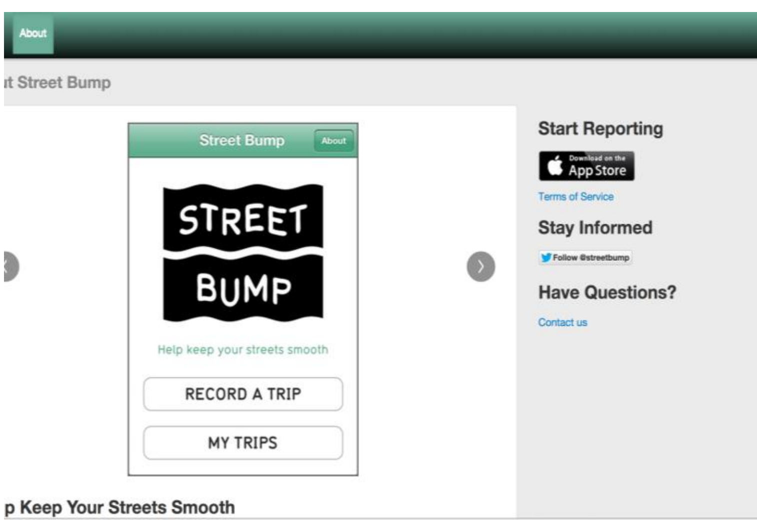
$P_i = \alpha + \beta_1 MoT_{i-2} + \beta_2 MoT_{i-4} + \beta_3 Gt_i + \beta_4 GT_{i-1} + \beta_5 GT_{i-3}$



Source: Vivi Yulaswati, Bappenas

Community-Based Platform to Track Damaged Roads in Yogyakarta

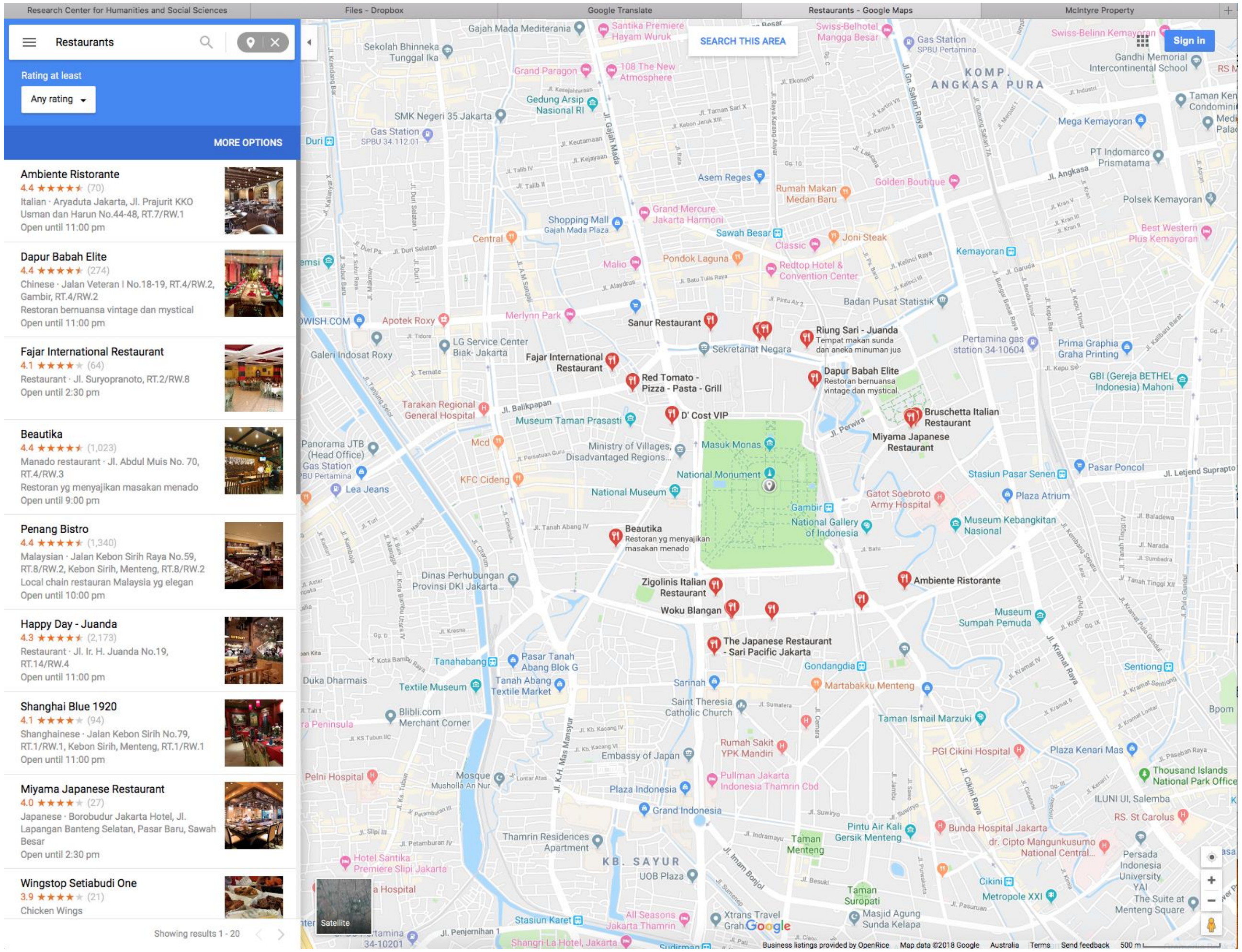




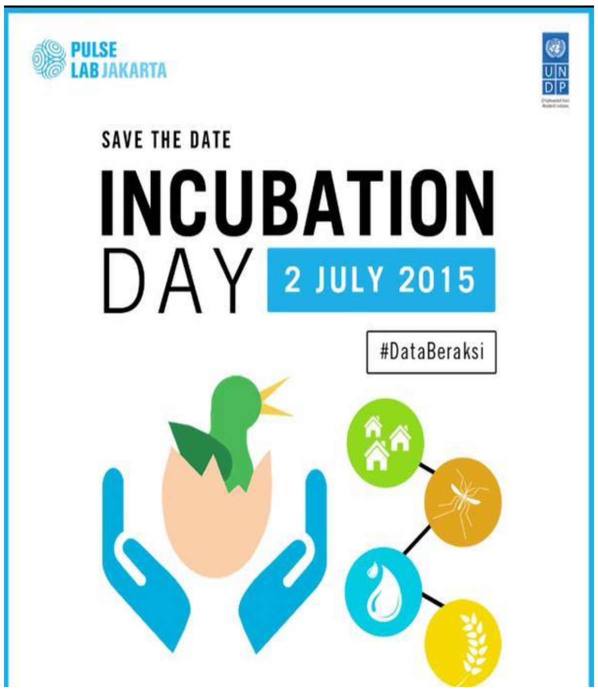
- JalananYogya is an approach to **leverage community participation** to report damaged roads in Yogyakarta, Indonesia
- Tool of data collection: using **geolocation feature** at **Twitter** or **mobile application**
- Aims:
  - To provide an efficient way of collecting data of damaged roads
  - To increase citizen’s awareness when driving
  - To reduce the number of accidents caused by the damaged road

Source: Vivi Yulaswati, Bappenas

Google map to understand the relation between education quality and the quality of restaurant



Data Pullers



Source: Vivi Yulaswati, Bappenas

Challenges

- Selection problem
- Measures collected may not be the best proxy of measures of interest
- Reliability issues
- Digital technology, data management and formatting
- Accessibility: data philanthropy, privacy concerns, proprietary data, security
- Analytical challenges: big data innovation and research

Australian National University

The way forward

- Providing more access to big data:
  - Gojek, Blue bird, climate remote sensing
  - Clear rules on using big data
- Technology (software) development for
  - Data mining and conversion
  - Analytic and data visualization
  - Storage and computation
- Expertise in using big data:
  - Collaboration in data mining; more systematic pullers
  - Research collaboration on using a specific big data for social analysis
  - Workshop and seminar utilizing big data