ECONOMETRICS
with
MACHINE LEARNING

Felix Chan and László Mátyás
Editors

Electronic Online Supplement

August 2022

Chapter 3
The Use of Machine Learning in Treatment Effect Estimation

1 The Derivation of Equation (3.13)

Let us abbreviate the notation for the estimation sample $\mathcal{S}^{est}$ to $\mathcal{S}$ and let X_t denote an observation on X which is independent of $\mathcal{S}$. Suppose that for any fixed subset $\ell \subset \mathcal{X}$, the estimator $\hat{\tau}_{\mathcal{S}}(\ell)$ is unbiased, i.e.,

$$E(\hat{\tau}_{\mathcal{S}}(\ell)) = \tau(\ell) \equiv E[Y(1) - Y(0) | X \in \ell]. \tag{1}$$

Recall that for a given partition $\Pi = \{\ell_1, \ldots, \ell_{\#\Pi}\}$ of $\mathcal{X}$, the CATE estimator is given by the step function

$$\hat{\tau}_{\mathcal{S}}(x; \Pi) = \sum_{j=1}^{\#\Pi} \hat{\tau}_{\mathcal{S}}(\ell_j) 1_{\ell_j}(x)$$

with expected mean squared error

$$\text{EMSE}(\Pi) = E_{X_t, \mathcal{S}}\left[\left(\tau(X_t) - \hat{\tau}_{\mathcal{S}}(X_t; \Pi)\right)^2\right].$$

We will show that

$$\text{EMSE}(\Pi) - E[\tau(X_t)^2] = E_{X_t}\left\{V_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(X_t; \Pi)]\right\} - E[\tau(X_t; \Pi)^2]. \tag{2}$$

Expanding the definition of EMSE, and using the law of iterated expectations, we can write

$$\begin{aligned}
&\text{EMSE}(\Pi) - E[\tau(X_t)^2] = E_{X_t, \mathcal{S}}\left\{\hat{\tau}^2_{\mathcal{S}}(X_t; \Pi) - 2\tau(X_t)\hat{\tau}_{\mathcal{S}}(X_t; \Pi)\right\} \\
&= E\left\{E_{X_t}\left[\hat{\tau}^2_{\mathcal{S}}(X_t; \Pi) - 2\tau(X_t)\hat{\tau}_{\mathcal{S}}(X_t; \Pi) \middle| \mathcal{S}, 1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right]\right\} \\
&= E\left\{\hat{\tau}^2_{\mathcal{S}}(X_t; \Pi) - 2\hat{\tau}_{\mathcal{S}}(X_t; \Pi) E_{X_t}\left[\tau(X_t) \middle| \mathcal{S}, 1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right]\right\} \qquad (3)
\end{aligned}$$

where the last equality uses the fact that conditional on $\mathcal{S}$ and $1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)$, the value of the step function estimator $\hat{\tau}_{\mathcal{S}}(X_t;\Pi)$ is pinned down (it is given by one of the values $\hat{\tau}_{\mathcal{S}}(\ell_j)$, $j = 1, \ldots, \#\Pi$). As X_t is independent of $\mathcal{S}$, we can further write

$$E_{X_t}\left[\tau(X_t)\middle|\mathcal{S}, 1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right] = E_{X_t}\left[\tau(X_t)\middle|1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right].$$

Now, by the law of iterated expectations, and the definition of $\tau(X_t)$,

$$E\left[\tau(X_t)\middle|1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right] = E\left[E\left[Y(1)-Y(0)|X_t\right]\middle|1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right]$$
$$= E\left[Y_t(1)-Y_t(0)\middle|1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right] = \sum_{j=1}^{\#\Pi}\tau(\ell_j)1_{\ell_j}(X_t),$$

because $E[Y_t(1)-Y_t(0)|\, X_t \in \ell_j] = \tau(\ell_j)$. Using the more compact notation

$$\tau(X_t;\Pi) = \sum_{j=1}^{\#\Pi}\tau(\ell_j)1_{\ell_j}(X_t),$$

we have shown that

$$E_{X_t}\left[\tau(X_t)\middle|\mathcal{S}, 1_{\ell_1}(X_t), \ldots, 1_{\ell_{\#\Pi}}(X_t)\right] = \tau(X_t;\Pi).$$

Substituting this result into (3) gives

$$\mathrm{EMSE}(\Pi) - E[\tau(X_t)^2] = E_{X_t,\mathcal{S}}\left\{\hat{\tau}^2_{\mathcal{S}}(X_t;\Pi) - 2\hat{\tau}_{\mathcal{S}}(X_t;\Pi)\tau(X_t;\Pi)\right\}$$
$$= E_{X_t,\mathcal{S}}\left\{\left[\hat{\tau}_{\mathcal{S}}(X_t;\Pi) - \tau(X_t;\Pi)\right]^2 - \tau^2(X_t;\Pi)\right\}$$
$$= E_{X_t}\left\{E_{\mathcal{S}}\left(\left[\hat{\tau}_{\mathcal{S}}(X_t;\Pi) - \tau(X_t;\Pi)\right]^2\middle|X_t\right)\right\} - E_{X_t}[\tau^2(X_t;\Pi)], \tag{4}$$

where the last equality again uses the law of iterated expectations. By the assumption that $\hat{\tau}_{\mathcal{S}}(\ell)$ is unbiased for $\tau(\ell)$, it follows immediately that

$$\tau(x;\Pi) = E_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(x;\Pi)]$$

for any fixed point $x \in \mathcal{X}$. Therefore,

$$E_{\mathcal{S}}\left(\left[\hat{\tau}_{\mathcal{S}}(x;\Pi) - \tau(x;\Pi)\right]^2\right) = V_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(x;\Pi)].$$

But because X_t is independent of $\mathcal{S}$,

$$E_{\mathcal{S}}\left(\left[\hat{\tau}_{\mathcal{S}}(X_t;\Pi) - \tau(X_t;\Pi)\right]^2\middle|X_t\right) = V_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(x;\Pi)]\big|_{x=X_t} = V_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(X_t;\Pi)].$$

Note that $V_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(X_t;\Pi)]$ is a random variable; it is the sampling variance of $\hat{\tau}_{\mathcal{S}}(x;\Pi)$ evaluated at the random point $x = X_t$. Substituting the previous result into (4) gives

$$\text{EMSE}(\Pi) - E[\tau(X_t)^2] = E_{X_t}\left\{V_{\mathcal{S}}[\hat{\tau}_{\mathcal{S}}(X_t;\Pi)]\right\} - E_{X_t}[\tau^2(X_t;\Pi)],$$

which is what we wanted to show.

2 Empirical Results for the White Subsample

These are the results for white mothers from the two exercises described in Section 3.5.

Table 1: Estimates of τ for white mothers

	Basic setup		Extended setup	
	Point-estimate	SE	Point-estimate	SE
OLS	-208.7966	2.4772	-207.5961	2.4952
Naive Lasso (λ^*)	-208.8705	-	-207.5191	-
Naive Lasso ($0.5\lambda^*$)	-208.9730	-	-208.6516	-
Naive Lasso ($2\lambda^*$)	-208.9030	-	-207.4272	-
Post-naive-lasso (λ^*)	-208.9526	2.4742	-206.988	2.4817
Post-naive-lasso ($0.5\lambda^*$)	-208.9526	2.4742	-205.7859	2.4736
Post-naive-lasso ($2\lambda^*$)	-208.7966	2.4772	-207.6697	2.4857
DML (λ^*)	-208.7972	2.4770	-206.3285	2.4895
DML ($0.5\lambda^*$)	-208.7802	2.4771	-205.9974	2.4903
DML ($2\lambda^*$)	-208.7740	2.4771	-206.4051	2.4895
DML-package	-208.7264	2.4772	-206.4665	2.4909

Notes: $N = 433,558$

Notes: Sample size= $433,558$. λ^* denotes lasso penalties obtained by 5-fold cross validation. The DML estimators are implemented by 2-fold cross-fitting. The row titled 'DML-package' contains the estimate obtained by using the 'official' DML code (`dml2`) available at https://docs.doubleml.org/r/stable/. All other estimators are programmed directly by the authors.

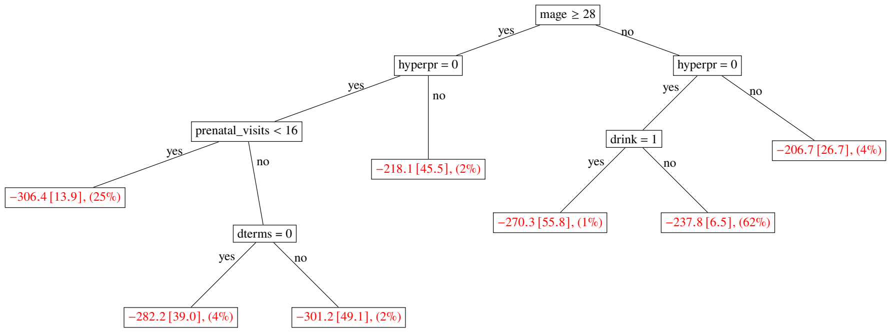

Fig. 1: A causal tree for the effect of smoking on birth weight (first time white mothers). Standard errors are in brackets and the percentages in parenthesis denote share of group in the sample. The total number of observations used to estimate the causal tree is a randomly drawn sample with $N = 150,000$, from the original $433,558$ sample. The covariates used are X_1, except for the polynomial and interaction terms. To obtain a simpler model, we choose the pruning parameter using the 1SE-rule rather than the minimum cross-validated MSE.

Chapter 9
Poverty, Inequality and Development Studies with Machine Learning

3 Contributions in Data Availability, Frequency and Granularity

Table 9.1 provides an exhaustive list of studies that aims at improving the availability, frequency and granularity of poverty, inequality, and development indicators. It describes the scope, the data sources and ML methods used and the contribution (whether it improves the frequency or granularity of estimates, if it covers rural areas, and if it includes visualizations) of each paper.

Table 2: Contributions in data availability, frequency and granularity

Paper	**What, Where**	**Data sources**	**Freq.**[1]	**Gran.**[1]	**Rural areas**	**Vis**[1]	**Relevant method**
Afzal et al. (2015)	Poverty, district level, Pakistan and Sri Lanka	Satellite imagery		X	X		Foward Stepwise, LASSO
Aiken et al. (2020)	Poverty, household level, Afghanistan	Mobile phone		X	X		Elastic Net, Decision Tree, Random Forest, XGBoost
Aiken et al. (2021)	Poverty, household level, rural Togo	Satellite imagery, Mobile phone		X	X		Gradient Boosting, PCA
Andreano et al. (2021)	Poverty, subnational level, Latin America	Satellite imagery	X	X	X	Vis in the paper	Panel models
Antenucci et al. (2014)	Job loss index US	Twitter	X		-	Vis in the paper	OLS

Zhongming et al. (2021)	Poverty, household level, rural Togo	Satellite imagery, Mobile phone		X	X		Gradient Boosting
Askitas and Zimmermann (2009)	Unemployment rate Germany	Google search data	X		-	Vis in the paper	Error Correction Model
Babenko et al. (2017)	Poverty, municipal level, Mexico	High and medium-resolution satellite imagery		X	X	Vis in the paper	CNN
Baylé (2016)	Slum detection, census segment level, municipality in Argentina	High-resolution satellite imagery, geo-located census, road and natural data from crowd-sourced maps		X	X	Vis in the paper	Random Forests, XGBoost, Support Vector Machines and Gaussian Mixtures

Blumenstock et al. (2018)	Wealth, individual level, Afghanistan	Mobile phone data		X	NS	Vis in the paper	Random Forests, XGBoost, Support Vector Machines and Gaussian Mixtures
Blumenstock et al. (2015)	Wealth, individual and cell level (smallest administrative unit), Rwanda	Mobile phone data	X	X	X	Vis in the paper	Feature engineering, Elastic Net, PCA
Bosco et al. (2017)	Socio-economic variables, 1x1 km resolution, 4 low-income countries	Geolocated household surveys		X	X	Vis in the paper	Integrated nested Laplace approximations and neural networks
Burke et al. (2016)	Under-5 mortality, 10kmx10km resolution, 28 sub-saharan African countries. 1980s-2000s	Geolocated Surveys (include location and timing of childbirths and deaths)	X	X	X	Vis in the paper	Kernel density estimator interpolation approaches, Linear Multivariate Regression
Chetty et al. (2018)	Children outcomes, census tract level, United States	Census, surveys and administrative data		X	X	Link	Lowess Regressions, OLS
Chi et al. (2022)	Wealth, 2.4x2.4 km resolution, 135 low and middle-income countries	Satellite imagery, Mobile phone, Social Media		X	X	Link	CNN, Principal Component Analysis, Gradient Boosting
Cuaresma et al. (2020)	Poverty rates, subnational regions, North Korea. 2012-2018	Remote-sensed night-time light intensity	X	X	X		KNN

Dahmani et al. (2014)	Slum detection, Moulay Bousselham (Morroco)	Very High-resolution satellite imagery		X	X	Vis in the paper	Support Vector Machines
Decuyper et al. (2014)	Poverty indices and food consumption, African country	Mobile phone data	X		X		OLS, Symbolic Regression
Doll et al. (2006)	GDP, 5 km resolution, 11 European countries and US	Satellite imagery		X	X	Vis in the paper	OLS
Duque et al. (2015)	Slum Index, 139 regions Medellin (Colombia)	Very High-resolution satellite imagery		X	no	Vis in the paper	Spatially Adjusted Regression and Stepwise Selection (Forward and Backward)
Eagle et al. (2010)	Development index, 32.482 communities UK	Mobile Phone data and Landlines (National communication network)		X	X	Vis in the paper	Network analysis, PCA
Ebener et al. (2005)	GPD, national and subnational level, 171 countries	Satellite imagery	X		-		OLS
Elbers et al. (2003)	Income, "town" (15000 households) level, Ecuador	Surveys, Census		X	X		ELL method
Ella et al. (2008)	Slum detection, Soweto (South Africa)	High-resolution imagery		X	no		Support Vector Machines
Elvidge et al. (2009)	Global poverty index, 30 arcsec resolution, national and subnational level	Satellite data (night-light), LandScan population counts	X	X	X	No longer available (Link)	-

Engstrom et al. (2017)	Poverty rates and average log consumption,1,291 administrative units, Sri Lanka	High-resolution satellite imagery		X	X		CNN and classification of spectral and textural characteristics
Ettredge et al. (2005)	Unemployment rate, US	Google search data	X		-		OLS
Farrell et al. (2020)	Gross family income, US. 2013-2017	Administrative Banking Data, Census	X		-		Gradient Boosting Machines (GBM)
Fatehkia et al. (2020)	Small area estimates Wealth Index, India and Philippines	Facebook data		X	X		Gradient Boosting Machines (GBM), Lasso Regression
Feldmeyer et al. (2020)	Socio-economic indicators, 1101 municipalities, Baden-Württemburg (Germany)	Crowsourced maps (OpenStreetMap)		X	X	Vis in the paper and github link provided upon completion of PhD	OLS, Random Forest, Neural Networks
Frias-Martinez and Virseda (2012)	Socio-economic indicators, national level, Latin America	Mobile phone data and Census	X		-		OLS

Gevaert et al. (2016)	Slum detection. 30x30m2 resolution, Kigaly (Rwanda)	High-resolution imagery from Unmanned Aerial Vehicles (UAVs)		X	X		Support Vector Machines with an RBF kernel (LibSVM), Local Binary Patterns (LBP), Mean Shift Algorithm, Spatial Binning, Correlation-Based Feature Selector
Ghosh et al. (2013)	Wellbeing indicators World map, 1km2 and subnational level	Night-time satellite imagery		X	X	Vis in the paper	OLS
Glaeser et al. (2018)	Income, block groups, New York	Google Street View imagery		X	no		Geometric Layout algorithm, V-Support Vector Regressor (v-SVR)
González-Fernández and González-Velasco (2018)	Unemployment rate, Spain	Google search data	X		-		OLS
Graesser et al. (2012)	Slum detection, 15x15m resolution, Caracas (Venezuela), La Paz (Bolivia), Kabul y Kandahar (Afganistán)	High-resolution satellite imagery		X	no	Vis in the paper	Decision Trees

Graetz et al. (2018)	Educational attainment, five km grids, Africa	Geolocated Surveys, Census		X	X	Vis in the paper . They can not share estimations but code for replication is available at: Link	Boosted Regression Trees and Lasso Regression
Head et al. (2017)	Wealth index, small regions, sub-Saharan Africa	High-resolution satellite imagery		X	X		CNN, Ridge Regression
Heitmann and Buri (2019)	Poverty, neighborhood levels, Gana and Uganda	High-resolution satellite imagery, mobile phone data		X	X	Vis in the paper	CNN, KNN spatial boosting
Henderson et al. (2012)	Income growth, national level, 188 countries	Satellite data	X		-		OLS
Hernandez et al. (2017)	Poverty rates, municipal level, Guatemala	Mobile phone data		X	X	Vis in the paper	Support Vector Machines, Random forests, Stochastic Gradient Boosting, K-means, Gaussian Mixtures, SuperVis in the papered Topic Models
Hersh et al. (2021)	Household labor income, district level, Belize	Open source satellite images		X	X	They can not share poverty estimations but code for replication is available. Link	Ridge Regression, Elastic Net, Random Forest, Extreme Gradient Boosting Trees

Hofer et al. (2020)	Poverty rate, 4km grid, Philippines and Thailand	Satellite imagery		X	X	Vis in the paper	CNN, Ridge Regression
Holzbauer et al. (2016)	GDP, state level, US	Social Media (Gowalla)	x		-		Linear models
Hristova et al. (2016)	Developement indexes, national level	Flow networks between countries (World Trade, Global migration, Digital communications, Flights)	X		-		Network analysis
Huang et al. (2015)	Slum detection, 120x120m, 2 mega cities in China	High-resolution satellite imagery		X	no		Random Forest, Support Vector Machines
Jean et al. (2016)	Poverty rate, village level, African countries (Nigeria, Tanzania, Uganda, Malawi y Rwanda)	High-resolution satellite imagery	X	X	X		Multistep transfer learning approach, CNN, PCA
Kavanagh et al. (2016)	Poverty Desentralization Index, region level, Scottish cities. 2001-2011	Census	X	X	no	Vis in the paper	Bayesian Multivariate Conditional Autorregresive (CAR) model
Khelifa and Mimoun (2012)	Slum detection, Oran (Algeria)	High-resolution satellite imagery		X	no		PCA, Genetic Algorithm
Lansley and Longley (2016)	Behaviour characteristics, London	Social media data (Twitter)		X	no	Vis in the paper	Latent Dirichlet Allocation (LDA)

Liu et al. (2016)	GDP, national level, China	Social network in China (Sina Microblog posts)	X		-	Vis in the paper	Support Vector Machines
Llorente et al. (2015)	Unemployment, 340 Spanish regions	Twitter data		X	X		OLS
Maiya and Babu (2018)	Slum detection, India	Satellite images		X	no		Mask-R-CNN
McBride and Nichols (2018)	Poverty rate, national level, East Timor, Malawi and Bolivia	Household surveys		X	-		Regression Forest, Quantile Reggresion Forest
McKenzie and Slind (2019)	Financial touch points, Kenia and Uganda	Facebook data, Twitter data, Open-Street-Map		X	X		OLS, Spatial Lag Regression, Support Vector Machines, Random Forest
Njuguna and McSharry (2017)	Multi-dimensional poverty index, sector level, Rwanda.	Mobile phone, satellite data (night lights), population density		X	X	Vis in the paper	Lasso Regression
Norbutas and Corten (2018)	Income per capita, 438 Ducth municipalities	Social media (Hyves)		X	X	Vis in the paper	OLS
Chen and Nordhaus (2015)	Population and output measures, African regions	Satellite data (night-time lights)		X	X		OLS

ILO-ECLAC (2018)	Child labor identification metodology, subnational level (lower level disaggregation possible), Latin America	Surveys and Census		X	X	Vis in the paper	Logistic Regression
Osgood-Zimmerman et al. (2018)	Malnutrition, 5x5 km resolution, 51 African countries. 2000-2015	Geolocated Surveys	X	X	X	Vis in the paper and code available: Link	Generalized Additive Models, Boosted Regression Trees, Lasso Regression, Bayesian Hierarchical Model
Otok and Seftiana (2014)	Classification of poor households, Jombang (Indonesia)	Survey		X	no	Vis in the paper	CART, Random Forest
Owen and Wong (2013)	Slum detection, Guatemala city	Very High Resolution imagery and Elevation data		X	no		CART, Discriminant Function Analysis
Perez et al. (2017)	Poverty rate, local level, Africa	Satellite imagery		X	X		CNN, Ridge Regression, Nearest-neighbor, Gradient Boosted Trees
Pokhriyal and Jacques (2017)	Multi-dimensional poverty index, Senegal	Mobile phone data, Enviromental data		X	X	Vis in the paper	Gaussian Process Regression, Elastic Net
Preis et al. (2012)	Future orientation index as proxy of GDP, 45 countries	Google search data	X		-		OLS

Quercia et al. (2012)	Gross Community Happiness, London	Social media data (Twitter)	x	X	no		NLP
Reiner Jr et al. (2018)	Childhood diarrheal morbidity and mortality, Africa, 2000–2015.	Geolocated Surveys	X	X	X	Vis in the paper	Spatial regression methods
Robinson et al. (2007)	Household expenditure, 0.01 degrees of resolution, Uganda	Very High Resolution imagery		X	X	Vis in the paper	Discriminant Analysis
Rogers et al. (2006)	Household expenditure, 0.01 degrees of resolution, Uganda	Remote-sensed satellite data, Survey		X	X	Vis in the paper	Discriminant Analysis
Rosati et al. (2020)	Health vulnerability, census block level, Argentina, 2010-2018	Census, Street grids	X	X	X	Link	Semiparametric PCA
Schmitt et al. (2018)	Slum detection, Cape Town (South Africa), Mumbai (India) and Manila (Philippines)	Very High Resolution imagery		X	no		Polarimetric Kennaugh, Schmittlets
Sheehan et al. (2019)	Asset Wealth and Education outcomes, subnational level, Africa	Wikipedia articles (geo-referenced)		X	X	Vis in the paper	NLP
Smith-Clarke et al. (2014)	Poverty rate and Assets Index, subnational regions, 2 developing countries	Mobile phone data		X	X	Vis in the paper	PCA
Sohnesen and Stender (2017)	Poverty rate, urban, rural and national level, six countries	Surveys			X		Random Forest, Lasso Regression, Stepwise selection

Soman et al. (2020)	Street-block accessibility index, worldwide	Crowsourced maps			no	Link	Topological Analysis
Steele et al. (2017)	Wealth Index, polygons up to 5 km, Bangladesh	Satellite, Mobile phone data		X	X	Vis in the paper	Bayesian Geoestatist-ical Models
Suraj et al. (2017)	Developement indicators, areas of 7km, India	High-resolution satellite imagery	X	X	X	Link	CNN
Tatem et al. (2014)	Poverty indices, 1x1 km pixel, Kenya, Uganda, Tanzania, Malawi, Nigeria and Pakistan	Geolocated Surveys, Satellite data		X	X	Vis in the paper	Model-based geostatistics
Tusting et al. (2019)	Housing conditions, 5x5 km resolution, Sub-Saharan Africa. 2000-2015	Geolocated Surveys	X	X	X	Vis in the paper	Geoestatistical regression model
Venerandi et al. (2015)	Index of multiple deprivation, small census areas, three urban zones UK	OpenStreetMap		X	no		Naive Bayes Classifier
Wang et al. (2019)	GDP, subnational level, China	Social Media (Weibo) and CVs from job seekers		X	NS		Naive Bayes Classifier
Watmough et al. (2019)	Wealth, household level, rural Kenya	Satellite sensor data		X	X		Classification tree
Weber et al. (2018)	Global Developement statistics, national level	Social media data (Face-book)	X		-	Vis in the paper	OLS
Weiss et al. (2018)	Global map of travel time to cities, 1 x 1km resolution	Open Street Map and Google		X	X	Link	Least-cost-path algorithm

Wurm et al. (2017)	Slum detection, Mumbai (India)	Remote-sensed data	X	no	Vis in the paper	Kennaugh matrix, Random Forest
Wurm et al. (2019)	Slum detection, Mumbai (India)	Very High-resolution satellite imagery	X	no		Fully CNN
Yeh et al. (2020)	Asset wealth, 20000 villages, 23 African countries	Multispectral high-resolution satellite imagery	X	no	Code to replicate GitHub Link	CNN
Zhao et al. (2019)	Multidimensional poverty index, 10 ×10 km resolution, Bangladesh	Satellite imagery, land cover map, road map and division headquarter location data	X	X	Vis in the paper	Random Forest

[1] "Gran." stands for "contributions in granularity", "Freq." stands for "contributions in frequency" and "Vis." stands for "interactive visuzalization publicly available".
Note: NS stands for Not specified, CNN stands for Convolutional neural network, NLP stands for Natural Language Processing.

4 ML Methods for Improving PID Measurements and Forecasts

Table 9.2 lists the methods used in the PID studies discussed in Section 9.2. (Measurement and Forecasting), i.e., studies that seek to improve the availability, frequency and granularity of PID estimators, as well as studies that seek to reduce dimensionality and deal with missing data. It includes the reviewed papers that used each method.

Chapter 1 of this books explains regularization methods; Chapter 2 and Chapter 6, network analysis; Chapter 4, boosting. Classic references explaining ML methods are Hastie et al. (2009), James et al. (2013), Murphy (2012), Bishop and Nasrabadi (2006); for natural language processing, Jurafsky and Martin (2014); for deep learning, Goodfellow et al. (2016).

Table 3: ML methods for improving PID measurements and forecasts - complete list of reviewed papers

Method		**Papers**
Supervised learning		
Trees and ensembles	Decision and regression trees	Graesser et al. (2012), Watmough et al. (2019), Aiken et al. (2020), Otok and Seftiana (2014)
	Boosting	Baylé (2016), Perez et al. (2017), Blumenstock et al. (2018), Graetz et al. (2018), Aiken et al. (2020), Farrell et al. (2020), Fatehkia et al. (2020), Aiken et al. (2021), Hersh et al. (2021), Osgood-Zimmerman et al. (2018), Chi et al. (2022), Hernandez et al. (2017), Heitmann and Buri (2019)
	Random forest	Thoplan (2014), Sohnesen and Stender (2017), Okiabera, Feigenbaum (2016),McBride and Nichols (2018), Huang et al. (2015), Baylé (2016), Hernandez et al. (2017), Wurm et al. (2017), McKenzie and Slind (2019), Zhao et al. (2019), Aiken et al. (2020), Feldmeyer et al. (2020), Hersh et al. (2021), Otok and Seftiana (2014)
Nonlinear regression methods	Generalized additive models	Osgood-Zimmerman et al. (2018), Bosco et al. (2017)
	Gaussian process regression	Pokhriyal and Jacques (2017), Tatem et al. (2014)
	Lowess regression	Chetty et al. (2018)

K nearest neighbors		Cuaresma et al. (2020), Perez et al. (2017), Heitmann and Buri (2019)
Naive Bayes		Venerandi et al. (2015)
Discriminant analysis		Owen and Wong (2013), Rogers et al. (2006), Pokhriyal and Jacques (2017), Robinson et al. (2007)
Support vector machines		Ella et al. (2008), Feigenbaum (2016), Huang et al. (2015), Baylé (2016), Gevaert et al. (2016), Liu et al. (2016), Glaeser et al. (2018), McKenzie and Slind (2019), Hernandez et al. (2017)
Regularization and feature selection	LASSO	Rosati (2017), Lucchetti (2018), Afzal et al. (2015), Njuguna and McSharry (2017), Sohnesen and Stender (2017), Graetz et al. (2018), Fatehkia et al. (2020), Osgood-Zimmerman et al. (2018), Lucchetti et al. (2018)
	Ridge regression	Jean et al. (2016),Perez et al. (2017), Hofer et al. (2020), Hersh et al. (2021), Head et al. (2017)
	Elastic Net	Doudchenko and Imbens (2016), Hersh et al. (2021), Clay et al. (2020), Aiken et al. (2020), Pokhriyal and Jacques (2017), Blumenstock et al. (2015)
	Wrapper feature selector	Mohamud and Gerek (2019), Afzal et al. (2015), Sohnesen and Stender (2017), Duque et al. (2015)

	Correlation feature selector	Gevaert et al. (2016)
Other spatial regression methods		Osgood-Zimmerman et al. (2018)
Deep learning	Neural networks	Jean et al. (2016), Dahmani et al. (2014), Babenko et al. (2017), Bosco et al. (2017), Engstrom et al. (2017), Perez et al. (2017), Wurm et al. (2017), Maiya and Babu (2018), Head et al. (2017), Suraj et al. (2017), Heitmann and Buri (2019), Hofer et al. (2020), Yeh et al. (2020), Chi et al. (2022), Rosati et al. (2020), Feldmeyer et al. (2020), Duque et al. (2015), Gevaert et al. (2016), Bosco et al. (2017), McKenzie and Slind (2019), Steele et al. (2017), Osgood-Zimmerman et al. (2018), Tusting et al. (2019),Kavanagh et al. (2016)
Unsupervised learning		
Factor analysis		Gasparini et al. (2013), Luzzi et al. (2008)
PCA (including its derivations, eg., sparse PCA)		Edo et al. (2021), Aiken et al. (2021), Duque et al. (2015), Rosati et al. (2020), Merola and Baulch (2019), Eagle et al. (2010), Chi et al. (2022), Blumenstock et al. (2015), Khelifa and Mimoun (2012), Jean et al. (2016), Smith-Clarke et al. (2014)
Clustering methods (e.g., k-means)		Caruso et al. (2015), Burke et al. (2016), Hernandez et al. (2017)

Processing new data	
Natural Language Processing	Quercia et al. (2012), Lansley and Longley (2016), Sheehan et al. (2019)
Other authomatized feature extraction (not deep learning)	Blumenstock et al. (2015), Glaeser et al. (2018), Graesser et al. (2012), Khelifa and Mimoun (2012), Gevaert et al. (2016)
Network analysis	Eagle et al. (2010), Hristova et al. (2016), Soto et al. (2011), UN Global (2016)
Traditional econometrics	
OLS	Liu et al. (2016), Ebener et al. (2005), Ettredge et al. (2005), Doll et al. (2006), Frias-Martinez and Virseda (2012), Preis et al. (2012), Henderson et al. (2012), Ghosh et al. (2013), Decuyper et al. (2014), Llorente et al. (2015), Chen and Nordhaus (2015), Norbutas and Corten (2018), González-Fernández and González-Velasco (2018), Weber et al. (2018), Chetty et al. (2018), Feldmeyer et al. (2020), Ghosh et al. (2013), McKenzie and Slind (2019), Burke et al. (2016)
Probit, Logit and derivations	Feigenbaum (2016), ILO-ECLAC (2018)
Panel models	Andreano et al. (2021)
PMM	Lucchetti et al. (2018)
Error-Correction Model	Askitas and Zimmermann (2009)

Other various methods	Elbers et al. (2003), Afzal et al. (2015), Schmitt et al. (2018), Weiss et al. (2018), Soman et al. (2020), Holzbauer et al. (2016), Edo et al. (2021), Decuyper et al. (2014)

5 Leveraging New Data Sources for Causal Inference

Table 9.3 lists PID studies that take advantage of the increased availability of new data sources for causal identification (see Section 9.3.5, New Data Sources for Outcomes and Treatments). It classifies the non-traditional data source use in three non-exclusive categories: outcome construction, treatment or control variable construction, and exogenous variability. It also describes the type of data source used, the evaluation method and, when applicable, the ML method employed, among other features of each paper.

Table 4: Papers leveraging new data sources for causal inference in PID studies

Paper	What, Where	New data source	New data use	Exogenous variability	Evaluation method	ML method
Alam et al. (2019)	Transport insfrastructure program, 16 countries	Satellite imagery	Outcome construction	Timing of the program	DiD	
Alix-Garcia et al. (2013)	CCT on environmental degradation, Mexico	Satellite imagery	Outcome construction	Experiment, discontinuity	Regression discontinuity	
Asfaw et al. (2017)	CCT and weather shocks on welfare, Zambia	Geo-referenced data	Treatment construction, control variables	Experiment, natural experiment	Quintile regression, GLS Random Effects	
BenYishay et al. (2017)	Land rights on deforestation, Brazil	Satellite imagery	Outcome construction	Timing of the program	DiD	

Dolan et al. (2019)	Anti-malarian program, Congo	Geo-referenced Survey	Outcome and control variables construction	Variation in malaria across subnational regions and timing of the program	DiD	
Blumenstock et al. (2015)	Mobile Salary Payment Program, Afghanistan	Mobile phone data	Outcome construction	Experiment	DiD	
Bunte et al. (2017)	Foreign direct investment, Liberia	Remotely sensed data (night time lights)	Outcome and controls construction	Variation of variables at subnational region level	Matching	
Chakravorty et al. (2016)	Impact of Electric access on welfare, Phillipines	Geo-referenced data	Exog. variability	Instrument	IV	
Chen et al. (2013)	Impact of Hua River Policy on life expectancy, China	Geo-referenced data	Outcome and treatment construction	Variation in air pollution	Regression discontinu-ity	
Chioda et al. (2016)	Impact of CCT on crime, Sao Paulo (Brazil)	Geo-referenced data	Outcome construction	Program design	IV, DiD	
Chor and Li (2021)	Impact of US-China tariff war on economic development, China	Satellite imagery (night time lights), Geo-referenced data	Outcome construction and exog. variability	Variation in exposure over time	IV, DiD	
Cohen et al. (2019)	Impact of closing airport on housing prices, Denver (US)	Web scrapping prices and location	Outcome construction		DiD	
Corbi et al. (2014)	Impact of federal transfers on local economic activity, Brazil	Satellite imagery (night time lights)	Outcome construction	Quasi-experimental policy variation	Fuzzy Regression discontinu-ity	Evaluation method

Dadvand et al. (2015)	Impact of green spaces on cognitive development, Barcelona (Spain)	Satellite imagery	Treatment construction	Variation of variables at subnational region level	Linear mixed effects model
Dakhlia et al. (2021)	Impact of finantial inclusion on economic development, Ethnic groups in Nigeria and Senegal	Satellite imagery (night time lights)	Outcome construction	Variation of variables at subnational region level	Fixed effect regression, Probit
De and Becker (2015)	Foreign aid (health, water and education), Malawi	Geo-referenced data	Treatment construction	Instrument, subnational treatment variation	IV, DiD, Matching
Dinkelman (2011)	Effects of Rural Electrification on Employment, South Africa	Geo-referenced data	Exog. variability	Instrument	IV
Duflo and Pande (2007)	Productivity and distributional effects of large irrigation dams, India	Geo-referenced data	Exog. variability	Instrument	IV
Ecker and Maystadt (2021)	Anti-poverty program and impact of civil war on child nutrition, Yemen	Geo-referenced data	Treatment construction	Variation in armed conflict intensity	DiD
Elliott et al. (2015)	Impact of typhoons on local economic activity, China	Satellite imagery (night time lights)	Outcome construction	Natural experiment	DiD
Faber and Gaubert (2019)	Effect of tourism on economic development, Mexico	Satellite imagery	Exog. variability	Instrument	IV
Ferraro and Simorangkir (2020)	Impact of anti-poverty program on deforestation, Indonesia	Satellite imagery	Outcome construction	Timing of the program	Difference in differences

Garcia et al. (2022)	Impact of short-term rentals on housing prices, Los Angeles (US)	Web scrapping from Airbnb	Outcome construction	Instrument	IV	
Hodler and Raschky (2014)	Regional favoritism from political leaders, 126 countries	Satellite imagery (night time lights)	Outcome construction	Variation of variables at subnational region level	Fixed effect regression	
Huang et al. (2021)	Anti-poverty program, Kenya	Satellite imagery	Outcome construction	Experiment	DiD	Outcome construc-tion. Mask R-CNN
Ismailov et al. (2019)	Mobile money use, Tanzania	Geo-referenced data, Satellite imagery (night time lights)	Exog. variability and Outcome construction	Instrument	IV	
Jedwab and Storeygard (2020)	Transport insfrastructure program, 39 countries in Africa	Geo-referenced data, Satellite imagery (night time lights)	Outcome and treatment construction	Instrument	IV	
Klomp (2016)	Impact of natural disasters on economic development, 140 countries	Satellite imagery (night time lights)	Outcome construction	Natural experiment	DiD	
Lipscomb et al. (2013)	Effects of Electrification on development, Brazil	Geo-referenced data	Treatment construction	Instrument	IV	
Manacorda and Tesei (2020)	Mobile phone coverage on mass political movilization, Africa	Geo-referenced data, open source data	Outcome and treatment construction	Natural experiment	IV	

Mensah (2021)	Impact of mobile phone use on economic development, 3419 subnational regions in 201 countries	Satellite imagery (night time lights), geo-referenced data	Exog. variability	Instrument	IV	
Michalopoulos and Papaionnau (2014)	Impact of national institutions on economic development, Africa	Satellite imagery (night time lights)	Outcome construction	Natural experiment	Spatial regression discontinuity, Matching	
Milesi et al. (2003)	Urban land development, 8 states from US	Satellite imagery	Outcome and treatment construction	Variation of variables at subnational region level	OLS	
Mutuku et al. (2011)	Impact of bednets on malaria transmission, Kenya	Satellite imagery, GPS data	Treatment and control variables construction	Experiment	Poisson regression	
Pierskalla and Hollenbach (2013)	Effect of mobile phone coverage on political violence, Africa	Geo-referenced data	Ouctome construction, Exog. variability	Instrument, Timing of coverage	OLS, DiD	
Ratledge et al. (2021)	Impact of Electric access on wealth, Uganda	Satellite imagery	Outcome construction	Timing of the program	Matrix completion, Synthetic control with elastic net	CNN model for outcome construction. Also ML for evaluation method
Russ et al. (2018)	Transport insfrastructure program, Nigeria	Geo-referenced data, Satellite imagery (night time lights)	Outcome construction, Exog. variability	Instrument	OLS, IV	

Salazar et al. (2021)	Technology adoption program, Dominican Republic	Satellite imagery	Outcome construction	Experiment	DiD, Event study	Continuous change detection and classification, Mann-Kendall to analyse trends
Smith and Wills (2018)	Impact of petroleo booms on Poverty and Inequality, 36 countries	High-resolution satellite imagery	Outcome construction	Size of the discovery and timing of price booms	DiD	
Hsiang and Jina (2014)	Effects of cyclones on economic growth, World	Satellite imagery	Treatment construction	Natural experiment	DiD	
Velilla and Bragança (2020)	Impact of commodities price shocks on local activity, Colombia	Satellite imagery (night time lights)	Outcome construction	Variation in intensity of coffee production and international coffee price	DiD	
Villa (2016)	Cash transfer program, Colombia	Satellite imagery (night time lights)	Outcome construction	Timing of the program	DiD	
Warr and Aung (2019)	Impact of cyclone on inequality between households, Myanmar	Satellite imagery	Treatment construction	Natural experiment	Fixed effect regression	
Witmer and O'Loughlin (2011)	Effects of wars on economic development, Regions from Russia and South Ossetia	Satellite imagery (night time lights)	Outcome construction	Variation in conflicts duration	OLS	
Zhang et al. (2007)	Impact of urban sprwal on soil resources, Nanjing (China)	Satellite imagery and soil map integrated by GIS	Outcome and treatment construction	Variation of variables at subnational region level	OLS	

Zou (2020)	Impact of short-term rentals on housing prices, Washington	Web scrapping from Airbnb	Outcome construction	Demand shock	First-difference model

Note: CCT stands for Conditional Cash Transfer, DID stands for Difference in differences, IV stands for Instrumental Variable

References

Afzal, M., Hersh, J. & Newhouse, D. (2015). Building a better model: Variable selection to predict poverty in pakistan and sri lanka. *World Bank Research Working Paper*.

Aiken, E., Bedoya, G., Coville, A. & Blumenstock, J. E. (2020). Targeting development aid with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan. In *Proceedings of the 3rd acm sigcas conference on computing and sustainable societies* (pp. 310–311).

Aiken, E., Bellue, S., Karlan, D., Udry, C. R. & Blumenstock, J. (2021). *Machine learning and mobile phone data can improve the targeting of humanitarian assistance* (Tech. Rep.). National Bureau of Economic Research.

Alam, M., Herrera Dappe, M., Melecky, M. & Goldblatt, R. (2019). Wider economic benefits of transport corridors. *Policy Research Working Paper World Bank*.

Alix-Garcia, J., McIntosh, C., Sims, K. R. & Welch, J. R. (2013). The ecological footprint of poverty alleviation: evidence from mexico's oportunidades program. *Review of Economics and Statistics*, *95*(2), 417–435.

Andreano, M. S., Benedetti, R., Piersimoni, F. & Savio, G. (2021). Mapping poverty of latin american and caribbean countries from heaven through night-light satellite images. *Social Indicators Research*, *156*(2), 533–562.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C. & Shapiro, M. D. (2014). *Using social media to measure labor market flows* (Tech. Rep.). National Bureau of Economic Research.

Asfaw, S., Carraro, A., Davis, B., Handa, S. & Seidenfeld, D. (2017). Cash transfer programmes, weather shocks and household welfare: evidence from a randomised experiment in zambia. *Journal of Development Effectiveness*, *9*(4), 419–442.

Askitas, N. & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*.

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A. & Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico. *arXiv preprint arXiv:1711.06323*.

Baylé, F. (2016). *Detección de villas y asentamientos informales en el partido de la matanza mediante teledetección y sistemas de información geográfica* (Unpublished doctoral dissertation). Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.

BenYishay, A., Heuser, S., Runfola, D. & Trichler, R. (2017). Indigenous land rights and deforestation: Evidence from the brazilian amazon. *Journal of Environmental Economics and Management*, *86*, 29–47.

Bishop, C. M. & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Blumenstock, J., Callen, M., Ghani, T. & Koepke, L. (2015). Promises and pitfalls of mobile money in afghanistan: evidence from a randomized control trial. In *Proceedings of the seventh international conference on information and communication technologies and development* (pp. 1–10).

Blumenstock, J. E. (2018). Estimating economic characteristics with phone data. In *Aea papers and proceedings* (Vol. 108, pp. 72–76).

Blumenstock, J. E., Cadamuro, G. & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A., ... others (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. *Journal of The Royal Society Interface*, *14*(129), 20160825.

Bunte, J. B., Desai, H., Gbala, K., Parks, B. & Runfola, D. M. (2017). Natural resource sector fdi and growth in post-conflict settings: Subnational evidence from liberia. *Williamsburg (VA): AidData*.

Burke, M., Heft-Neal, S. & Bendavid, E. (2016). Sources of variation in under-5 mortality across sub-saharan africa: a spatial analysis. *The Lancet Global Health*, *4*(12), e936–e945.

Caruso, G., Sosa-Escudero, W. & Svarc, M. (2015). Deprivation and the dimensionality of welfare: a variable-selection cluster-analysis approach. *Review of Income and Wealth*, *61*(4), 702–722.

Chakravorty, U., Emerick, K. & Ravago, M.-L. (2016). Lighting up the last mile: The benefits and costs of extending electricity to the rural poor. *Resources for the Future Discussion Paper*, 16–22.

Chen, X. & Nordhaus, W. (2015). A test of the new viirs lights data set: Population and economic output in africa. *Remote Sensing*, *7*(4), 4937–4947.

Chen, Y., Ebenstein, A., Greenstone, M. & Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. *Proceedings of the National Academy of Sciences*, *110*(32), 12936–12941.

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. & Porter, S. R. (2018). *The opportunity atlas: Mapping the childhood roots of social mobility* (Tech. Rep.). National Bureau of Economic Research.

Chi, G., Fang, H., Chatterjee, S. & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, *119*(3).

Chioda, L., De Mello, J. M. & Soares, R. R. (2016). Spillovers from conditional cash transfer programs: Bolsa família and crime in urban brazil. *Economics of Education Review*, *54*, 306–320.

Chor, D. & Li, B. (2021). *Illuminating the effects of the us-china tariff war on china's economy* (Tech. Rep.). National Bureau of Economic Research.

Clay, K., Egedesø, P. J., Hansen, C. W., Jensen, P. S. & Calkins, A. (2020). Controlling tuberculosis? evidence from the first community-wide health experiment. *Journal of Development Economics*, *146*, 102510.

Cohen, J., Coughlin, C. C., Crews, J. & Ross, S. L. (2019). Negative externalities and real asset prices: Closing of stapleton airport and effect on nearby housing markets. *Available at SSRN 3473135*.

Corbi, R., Papaioannou, E. & Surico, P. (2014). *Federal transfer multipliers: Quasi-experimental evidence from brazil*. Citeseer.

Cuaresma, J. C., Danylo, O., Fritz, S., Hofer, M., Kharas, H. & Bayas, J. C. L. (2020). What do we know about poverty in north korea? *Palgrave Communications*, *6*(1), 1–8.

Dadvand, P., Nieuwenhuijsen, M. J., Esnaola, M., Forns, J., Basagaña, X., Alvarez-Pedrerol, M., . . . others (2015). Green spaces and cognitive development in primary schoolchildren. *Proceedings of the National Academy of Sciences*, *112*(26), 7937–7942.

Dahmani, R., Fora, A. A. & Sbihi, A. (2014). Extracting slums from high-resolution satellite images. *Int. J. Eng. Res. Dev*, *10*, 1–10.

Dakhlia, S., Diallo, B. & Temimi, A. (2021). Financial inclusion and ethnic development: Evidence from satellite light density at night. *Journal of Behavioral and Experimental Finance*, *29*, 100455.

De, R. & Becker, C. (2015). The foreign aid effectiveness debate: Evidence from malawi. *Online], vol. March, no. Working Paper*, *6*.

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J.-M., Krings, G., Gutierrez, T., . . . Luengo-Oroz, M. A. (2014). Estimating food consumption and poverty indices with mobile phone data. *arXiv preprint arXiv:1412.2595*.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from south africa. *American Economic Review*, *101*(7), 3078–3108.

Dolan, C. B., BenYishay, A., Grépin, K. A., Tanner, J. C., Kimmel, A. D., Wheeler, D. C. & McCord, G. C. (2019). The impact of an insecticide treated bednet campaign on all-cause child mortality: A geospatial impact evaluation from the democratic republic of congo. *PloS one*, *14*(2), e0212890.

Doll, C. N., Muller, J.-P. & Morley, J. G. (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, *57*(1), 75–92.

Doudchenko, N. & Imbens, G. W. (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis* (Tech. Rep.). National Bureau of Economic Research.

Duflo, E. & Pande, R. (2007). Dams. *The Quarterly Journal of Economics*, *122*(2), 601–646.

Duque, J. C., Patino, J. E., Ruiz, L. A. & Pardo-Pascual, J. E. (2015). Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data. *Landscape and Urban Planning*, *135*, 11–21.

Eagle, N., Macy, M. & Claxton, R. (2010). Network diversity and economic development. *Science*, *328*(5981), 1029–1031.

Ebener, S., Murray, C., Tandon, A. & Elvidge, C. C. (2005). From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. *international Journal of health geographics*, *4*(1), 1–17.

Ecker, O. & Maystadt, J.-F. (2021). Civil conflict, cash transfers, and child nutrition in yemen. *Households in Conflict Network*.

Edo, M., Escudero, W. S. & Svarc, M. (2021). A multidimensional approach to measuring the middle class. *The Journal of Economic Inequality*, *19*(1), 139–162.

Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*(1), 355–364.

Ella, L. A., van den Bergh, F., van Wyk, B. J. & van Wyk, M. A. (2008). A comparison of texture feature algorithms for urban settlement classification. In *Igarss 2008-2008 ieee international geoscience and remote sensing symposium* (Vol. 3, pp. III–1308).

Elliott, R. J., Strobl, E. & Sun, P. (2015). The local impact of typhoons on economic activity in china: A view from outer space. *Journal of Urban Economics*, *88*, 50–66.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B. & Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences*, *35*(8), 1652–1660.

Engstrom, R., Hersh, J. S. & Newhouse, D. L. (2017). Poverty from space: using high-resolution satellite imagery for estimating economic well-being. *World Bank Policy Research Working Paper*(8284).

Ettredge, M., Gerdes, J. & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, *48*(11), 87–92.

Faber, B. & Gaubert, C. (2019). Tourism and economic development: Evidence from mexico's coastline. *American Economic Review*, *109*(6), 2245–93.

Farrell, D., Greig, F. & Deadman, E. (2020). Estimating family income from administrative banking data: A machine learning approach. In *Aea papers and proceedings* (Vol. 110, pp. 36–41).

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M. & Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, *9*(1), 22.

Feigenbaum, J. J. (2016). A machine learning approach to census record linking. *Retrieved March*, *28*, 2016.

Feldmeyer, D., Meisch, C., Sauter, H. & Birkmann, J. (2020). Using openstreetmap data and machine learning to generate socio-economic indicators. *ISPRS International Journal of Geo-Information*, *9*(9), 498.

Ferraro, P. J. & Simorangkir, R. (2020). Conditional cash transfers to alleviate poverty also reduced deforestation in indonesia. *Science Advances*, *6*(24),

eaaz1298.
Frias-Martinez, V. & Virseda, J. (2012). On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development* (pp. 76–84).
Garcia, B., Miller, K. & Morehouse, J. (2022). In search of peace and quiet: The heterogeneous impacts of short-term rentals on housing prices. *The Center for Growth and Opportunity*.
Gasparini, L., Sosa-Escudero, W., Marchionni, M. & Olivieri, S. (2013). Multidimensional poverty in latin america and the caribbean: new evidence from the gallup world poll. *The Journal of Economic Inequality*, *11*(2), 195–214.
Gevaert, C., Persello, C., Sliuzas, R. & Vosselman, G. (2016). Classification of informal settlements through the integration of 2d and 3d features extracted from uav data. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, *3*, 317.
Ghosh, T., Anderson, S. J., Elvidge, C. D. & Sutton, P. C. (2013). Using nighttime satellite imagery as a proxy measure of human well-being. *sustainability*, *5*(12), 4988–5019.
Glaeser, E. L., Kominers, S. D., Luca, M. & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, *56*(1), 114–137.
González-Fernández, M. & González-Velasco, C. (2018). Can google econometrics predict unemployment? evidence from spain. *Economics Letters*, *170*, 42–45.
Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT Press. (http://www.deeplearningbook.org)
Graesser, J., Cheriyadat, A., Vatsavai, R. R., Chandola, V., Long, J. & Bright, E. (2012). Image based characterization of formal and informal neighborhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *5*(4), 1164–1176.
Graetz, N., Friedman, J., Osgood-Zimmerman, A., Burstein, R., Biehl, M. H., Shields, C., . . . others (2018). Mapping local variation in educational attainment across africa. *Nature*, *555*(7694), 48–53.
Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.
Head, A., Manguin, M., Tran, N. & Blumenstock, J. E. (2017). Can human development be measured with satellite imagery? In *Ictd* (pp. 8–1).
Heitmann, S. & Buri, S. (2019). Poverty estimation with satellite imagery at neighborhood levels.
Henderson, J. V., Storeygard, A. & Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, *102*(2), 994–1028.
Hernandez, M., Hong, L., Frias-Martinez, V., Whitby, A. & Frias-Martinez, E. (2017). Estimating poverty using cell phone data: evidence from guatemala. *World Bank Policy Research Working Paper*(7969).
Hersh, J., Engstrom, R. & Mann, M. (2021). Open data for algorithms: mapping poverty in belize using open satellite derived features and machine learning.

Information Technology for Development, *27*(2), 263–292.

Hodler, R. & Raschky, P. A. (2014). Regional favoritism. *The Quarterly Journal of Economics*, *129*(2), 995–1033.

Hofer, M., Sako, T., Martinez Jr, A., Addawe, M., Bulan, J., Durante, R. L. & Martillan, M. (2020). Applying artificial intelligence on satellite imagery to compile granular poverty statistics. *Asian Development Bank Economics Working Paper Series*(629).

Holzbauer, B. O., Szymanski, B. K., Nguyen, T. & Pentland, A. (2016). Social ties as predictors of economic development. In *International conference and school on network science* (pp. 178–185).

Hristova, D., Rutherford, A., Anson, J., Luengo-Oroz, M. & Mascolo, C. (2016). The international postal network and other global flows as proxies for national wellbeing. *PloS one*, *11*(6), e0155976.

Hsiang, S. M. & Jina, A. S. (2014). *The causal effect of environmental catastrophe on long-run economic growth: Evidence from 6,700 cyclones* (Tech. Rep.). National Bureau of Economic Research.

Huang, L. Y., Hsiang, S. & Gonzalez-Navarro, M. (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. *arXiv preprint arXiv:2104.11772*.

Huang, X., Liu, H. & Zhang, L. (2015). Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3639–3657.

ILO-ECLAC. (2018). *Child labour risk identification model. methodology to design preventive strategies at local level.* https://dds.cepal.org/redesoc/publicacion?id=4886.

Ismailov, A., Kimaro, A. B., Naito, H. et al. (2019). The effect of mobile money usage on borrowing, saving, and receiving remittances: Evidence from tanzania. *University of Tsukuba*, 2019–002.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.

Jedwab, R. & Storeygard, A. (2020). The average and heterogeneous effects of transportation investments: Evidence from sub-saharan africa 1960-2010. *Journal of the European Economic Association*.

Jurafsky, D. & Martin, J. H. (2014). Speech and language processing. vol. 3. *US: Prentice Hall*.

Kavanagh, L., Lee, D. & Pryce, G. (2016). Is poverty decentralizing? quantifying uncertainty in the decentralization of urban poverty. *Annals of the American Association of Geographers*, *106*(6), 1286–1298.

Khelifa, D. & Mimoun, M. (2012). Object-based image analysis and data mining for building ontology of informal urban settlements. In *Image and signal processing for remote sensing xviii* (Vol. 8537, p. 85371I).

Klomp, J. (2016). Economic development and natural disasters: A satellite data analysis. *Global Environmental Change*, *36*, 67–88.

Lansley, G. & Longley, P. A. (2016). The geography of twitter topics in london. *Computers, Environment and Urban Systems*, *58*, 85–96.

Lipscomb, M., Mobarak, A. M. & Barham, T. (2013). Development effects of electrification: Evidence from the topographic placement of hydropower plants in brazil. *American Economic Journal: Applied Economics*, *5*(2), 200–231.

Liu, J.-H., Wang, J., Shao, J. & Zhou, T. (2016). Online social activity reflects economic status. *Physica A: Statistical Mechanics and its Applications*, *457*, 581–589.

Llorente, A., Garcia-Herranz, M., Cebrian, M. & Moro, E. (2015). Social media fingerprints of unemployment. *PloS one*, *10*(5), e0128692.

Lucchetti, L. (2018). What can we (machine) learn about welfare dynamics from cross-sectional data? *World Bank Policy Research Working Paper*(8545).

Lucchetti, L., Corral, P., Ham, A. & Garriga, S. (2018). Lassoing welfare dynamics with cross-sectional data.

Luzzi, G. F., Flückiger, Y. & Weber, S. (2008). A cluster analysis of multidimensional poverty in switzerland. In *Quantitative approaches to multidimensional poverty measurement* (pp. 63–79). Springer.

Maiya, S. R. & Babu, S. C. (2018). Slum segmentation and change detection: A deep learning approach. *arXiv preprint arXiv:1811.07896*.

Manacorda, M. & Tesei, A. (2020). Liberation technology: Mobile phones and political mobilization in africa. *Econometrica*, *88*(2), 533–567.

McBride, L. & Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, *32*(3), 531–550.

McKenzie, G. & Slind, R. T. (2019). A user-generated data based approach to enhancing location prediction of financial services in sub-saharan africa. *Applied geography*, *105*, 25–36.

Mensah, J. T. (2021). Mobile phones and local economic development: A global evidence. *Available at SSRN 3811765*.

Merola, G. M. & Baulch, B. (2019). Using sparse categorical principal components to estimate asset indices: new methods with an application to rural southeast asia. *Review of Development Economics*, *23*(2), 640–662.

Michalopoulos, S. & Papaioannou, E. (2014). National institutions and subnational development in africa. *The Quarterly journal of economics*, *129*(1), 151–213.

Milesi, C., Elvidge, C. D., Nemani, R. R. & Running, S. W. (2003). Assessing the impact of urban land development on net primary productivity in the southeastern united states. *Remote Sensing of Environment*, *86*(3), 401–410.

Mohamud, J. H. & Gerek, O. N. (2019). Poverty level characterization via feature selection and machine learning. In *2019 27th signal processing and communications applications conference (siu)* (pp. 1–4).

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Mutuku, F. M., King, C. H., Mungai, P., Mbogo, C., Mwangangi, J., Muchiri, E. M., ... Kitron, U. (2011). Impact of insecticide-treated bed nets on malaria transmission indices on the south coast of kenya. *Malaria journal*, *10*(1), 1–14.

Njuguna, C. & McSharry, P. (2017). Constructing spatiotemporal poverty indices from big data. *Journal of Business Research*, *70*, 318–327.

Norbutas, L. & Corten, R. (2018). Network structure and economic prosperity in municipalities: A large-scale test of social capital theory using social media data. *Social networks*, *52*, 120–134.

Okiabera, J. O. (2020). *Using random forest (rf) to identify key determinants of poverty in kenya.* (Unpublished doctoral dissertation). University of Nairobi.

Osgood-Zimmerman, A., Millear, A. I., Stubbs, R. W., Shields, C., Pickering, B. V., Earl, L., . . . others (2018). Mapping child growth failure in africa between 2000 and 2015. *Nature*, *555*(7694), 41–47.

Otok, B. W. & Seftiana, D. (2014). The classification of poor households in jombang with random forest classification and regression trees (rf-cart) approach as the solution in achieving the 2015 indonesian mdgs' targets. *International Journal of Science and Research (IJSR) Volume*, *3*.

Owen, K. K. & Wong, D. W. (2013). An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics. *Applied Geography*, *38*, 107–118.

Perez, A., Yeh, C., Azzari, G., Burke, M., Lobell, D. & Ermon, S. (2017). Poverty prediction with public landsat 7 satellite imagery and machine learning. *arXiv preprint arXiv:1711.03654*.

Pierskalla, J. H. & Hollenbach, F. M. (2013). Technology and collective action: The effect of cell phone coverage on political violence in africa. *American Political Science Review*, *107*(2), 207–224.

Pokhriyal, N. & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, *114*(46), E9783–E9792.

Preis, T., Moat, H. S., Stanley, H. E. & Bishop, S. R. (2012). Quantifying the advantage of looking forward. *Scientific reports*, *2*(1), 1–2.

Quercia, D., Ellis, J., Capra, L. & Crowcroft, J. (2012). Tracking" gross community happiness" from tweets. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 965–968).

Ratledge, N., Cadamuro, G., De la Cuesta, B., Stigler, M. & Burke, M. (2021). *Using satellite imagery and machine learning to estimate the livelihood impact of electricity access* (Tech. Rep.). National Bureau of Economic Research.

Reiner Jr, R. C., Graetz, N., Casey, D. C., Troeger, C., Garcia, G. M., Mosser, J. F., . . . others (2018). Variation in childhood diarrheal morbidity and mortality in africa, 2000–2015. *New England Journal of Medicine*, *379*(12), 1128–1138.

Robinson, T., Emwanu, T. & Rogers, D. (2007). Environmental approaches to poverty mapping: an example from uganda. *Information development*, *23*(2-3), 205–215.

Rogers, D., Emwanu, T. & Robinson, T. (2006). Poverty mapping in uganda: an analysis using remotely sensed and other environmental data.

Rosati, G. (2017). Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning: Aplicación en la encuesta permanente de hogares (eph). *SaberEs*, *9*(1), 91–111.

Rosati, G., Olego, T. A. & Vazquez Brust, H. A. (2020). Building a sanitary vulnerability map from open source data in argentina (2010-2018). *International Journal for Equity in Health*, *19*(1), 1–16.

Russ, J., Berg, C., Damania, R., Barra, A. F., Ali, R. & Nash, J. (2018). Evaluating transport infrastructure projects in low data environments: an application to nigeria. *The Journal of Development Studies*, *54*(8), 1406–1425.

Salazar, L., Palacios, A., Selvaraj, M. & Montenegro, F. (2021). Using satellite images to measure crop productivity: Long-term impact assessment of a randomized technology adoption program in the dominican republic.

Schmitt, A., Sieg, T., Wurm, M. & Taubenböck, H. (2018). Investigation on the separability of slums by multi-aspect terrasar-x dual-co-polarized high resolution spotlight images based on the multi-scale evaluation of local distributions. *International journal of applied earth observation and geoinformation*, *64*, 181–198.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., . . . Ermon, S. (2019). Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2698–2706).

Smith, B. & Wills, S. (2018). Left in the dark? oil and rural poverty. *Journal of the Association of Environmental and Resource Economists*, *5*(4), 865–904.

Smith-Clarke, C., Mashhadi, A. & Capra, L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 511–520).

Sohnesen, T. P. & Stender, N. (2017). Is random forest a superior methodology for predicting poverty? an empirical assessment. *Poverty & Public Policy*, *9*(1), 118–133.

Soman, S., Beukes, A., Nederhood, C., Marchio, N. & Bettencourt, L. (2020). Worldwide detection of informal settlements via topological analysis of crowdsourced digital maps. *ISPRS International Journal of Geo-Information*, *9*(11), 685.

Soto, V., Frias-Martinez, V., Virseda, J. & Frias-Martinez, E. (2011). Prediction of socioeconomic levels using cell phone records. In *International conference on user modeling, adaptation, and personalization* (pp. 377–388).

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., . . . others (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, *14*(127), 20160690.

Suraj, P. K., Gupta, A., Sharma, M., Paul, S. B. & Banerjee, S. (2017). On monitoring development indicators using high resolution satellite images. *arXiv preprint arXiv:1712.02282*.

Tatem, A., Gething, P., Pezzulo, C., Weiss, D. & Bhatt, S. (2014). Final report: Development of high-resolution gridded poverty surfaces. *United Kingdom: University of Southampton University of Oxford*, *10*.

Thoplan, R. (2014). Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research (IJSBAR), North America*, *17*.

Tusting, L. S., Bisanzio, D., Alabaster, G., Cameron, E., Cibulskis, R., Davies, M., . . . others (2019). Mapping changes in housing in sub-saharan africa from 2000 to 2015. *Nature*, *568*(7752), 391–394.

UN Global. (2016). *Building proxy indicators of national wellbeing with postal data.* Project Series, no. 22, https://www.unglobalpulse.org/document/building-proxy-indicators-of-national-wellbeing-with-postal-data/.

Velilla, R. & Bragança, A. (2020). Coffee price shock and local economic performance: Evidence from colombia.

Venerandi, A., Quattrone, G., Capra, L., Quercia, D. & Saez-Trumper, D. (2015). Measuring urban deprivation from user generated content. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing* (pp. 254–264).

Villa, J. M. (2016). Social transfers and growth: Evidence from luminosity data. *Economic Development and Cultural Change*, *65*(1), 39–61.

Wang, J., Gao, J., Liu, J.-H., Yang, D. & Zhou, T. (2019). Regional economic status inference from information flow and talent mobility. *EPL (Europhysics Letters)*, *125*(6), 68002.

Warr, P. & Aung, L. L. (2019). Poverty and inequality impact of a natural disaster: Myanmar's 2008 cyclone nargis. *World Development*, *122*, 446–461.

Watmough, G. R., Marcinko, C. L., Sullivan, C., Tschirhart, K., Mutuo, P. K., Palm, C. A. & Svenning, J.-C. (2019). Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences*, *116*(4), 1213–1218.

Weber, I., Kashyap, R. & Zagheni, E. (2018). Using advertising audience estimates to improve global development statistics. *Itu Journal: Ict Discoveries*, *1*(2).

Weiss, D. J., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A., . . . others (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, *553*(7688), 333–336.

Witmer, F. D. & O'Loughlin, J. (2011). Detecting the effects of wars in the caucasus regions of russia and georgia using radiometrically normalized dmsp-ols nighttime lights imagery. *GIScience & Remote Sensing*, *48*(4), 478–500.

Wurm, M., Stark, T., Zhu, X. X., Weigand, M. & Taubenböck, H. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, *150*, 59–69.

Wurm, M., Taubenböck, H., Weigand, M. & Schmitt, A. (2017). Slum mapping in polarimetric sar data using spatial features. *Remote sensing of environment*, *194*, 190–204.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., . . . Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, *11*(1), 1–11.

Zhang, X., Chen, J., Tan, M. & Sun, Y. (2007). Assessing the impact of urban sprawl on soil resources of nanjing city using satellite images and digital soil databases. *Catena*, *69*(1), 16–30.

Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C. & Wu, J. (2019). Estimation of poverty using random forest regression with multi-source data: A case study in bangladesh. *Remote Sensing*, *11*(4), 375.

Zhongming, Z., Linong, L., Wangqiang, Z., Wei, L. et al. (2021). Mapping the spatial distribution of poverty using satellite imagery in thailand.

Zou, Z. (2020). Examining the impact of short-term rentals on housing prices in washington, dc: Implications for housing policy and equity. *Housing Policy Debate*, *30*(2), 269–290.