Using Data Differently and Using Different Data

The lack of adequate measures is often an impediment to robust policy evaluation. We discuss three approaches to measurement and data usage that have the potential to improve the way we conduct impact evaluations. First, the creation of new measures, when no adequate ones are available. Second, the use of multiple measures when a single one is not appropriate. And third, the use of machine learning algorithms to evaluate and understand programme impacts. We motivate the relevance of each of the categories by providing examples where they have proved useful in the past. We discuss the challenges and risks involved in each strategy and conclude with an outline of promising directions for future work.

Keywords: data; measurement; impact evaluation, machine learning

Impact evaluations, data and measurement

Poverty alleviation policies, in developing countries and elsewhere, are designed by an accumulation of evidence that directs efforts towards the most effective interventions. Impact evaluations are a key input into this process. These evaluations require data that measure outcomes as precisely as possible, to assess whether a given project has worked or not. In some cases, these outcomes can be observed and recorded easily. For example, take the case of a policy aimed at increasing school attendance. Attendance rates are generally well recorded, and if available to researchers, provide a suitable measure with which to measure the success of said policy. In other contexts, designing and collecting these measures may prove more challenging. Examples of this last case include measuring childhood development, risk aversion or expectations about health benefits and change in social norms.

To maximise the insight provided by an evaluation, researchers will also be interested in understanding the channels of impact, if any. This may help uncover the constraints under which households and individuals operate. As a result, a clearer

understanding of the conditions under which policies are (and are not) effective may emerge. The estimation of heterogeneous treatment effects, mediation analysis and structural models, are alternative paths for exploring these questions. This task requires more than just outcome measures. Ideally, all possible factors that could play a role in enabling or constraining a programme's effectiveness must be measured as well. Soon, the data and measurement requirements for an accurate and insightful evaluation become a significant challenge.

As a result, impact evaluations are often impeded by poor quality data. This could be due to missing variables of interest, mis-measured variables, or non-representative samples, for example. A large fraction of the literature on impact evaluations deals with research design. Indeed, randomised controlled trials (RCTs) for example, allow researchers to overcome one source of measurement concerns: the omission of unmeasured potential outcomes. In other words, since we cannot observe how individuals perform both with and without a given treatment, RCTs provide us with two samples that are identical, on average, which can be used to infer causal impacts. However, RCTs do not help with: (a) adequately measuring outcomes of interest; (b) measuring factors that relate to the mechanisms of estimated impacts; (c) conducting robust evaluations in areas where the RCT methodology is not feasible; and (d) ensuring the study setting/population is representative of the intended targets. Innovation in the field of measurement that capitalizes on the glut of new data available is critical to achieve both rigorous evaluations, and to increase the breadth of questions that can be addressed.

In this paper we discuss three approaches to measurement and data usage that have the potential to improve the way we conduct impact evaluations. We motivate the relevance of each of the categories by providing an overview of the key issues, using a

few detailed examples and concluding with an outline of promising directions for future work.

We start in Section 2 with a discussion of recent advances in the construction of new measures. Such measures would not only improve the precision of the impact evaluation but would also allow researchers to shed light on the mechanisms behind any detected impacts. The challenge is to develop measures that address the core parameters of an underlying theoretical model and meet the empirical requirements for the application of appropriate estimation methods.

The second category of approaches relates to the use of multiple measures. Even in areas where measures do exist they often only partially capture the underlying construct of interest. For example, cognitive outcomes are measured using a battery of tests, as opposed to a single one, in the understanding that each individual test is not a sufficient measure if taken on its own. A second-best approach in the face of imperfect measurement is therefore to use multiple incomplete/flawed measures, and combine them in an index that hopefully captures the underlying characteristic being estimated. Section 3 describes the reasons for this approach, and highlights some important pitfalls in the construction of these indices.

Finally, we discuss ways of leveraging large datasets and novel machine learning (ML) algorithms to improve the quality and reliability of impact evaluations. The advent of Big Data made large amounts of data available for researchers at high degrees of resolution. At the same time high, affordable, processing power made the handling of large amounts of data easier. This opened up the possibility of developing new measures built using these data, the combination of different sources of data, and the innovative use of high-dimensional datasets and ML algorithms. In Section 4, we focus on specific tasks, common to most impact evaluations, in which ML methods

have been proposed as a principled way to, for example, select regression models, identify the best set of instruments, or explore the heterogeneous impacts of a given policy.

Creating new measures

A significant fraction of the impact evaluation literature addresses the fundamental problem of causal inference: the fact that researchers only ever observe one potential outcome for each unit studied. A second, and arguably equally important challenge, is that of the appropriate measurement of outcomes and other underlying characteristics of the individuals in the study.

In some contexts, the participants' outcomes and characteristics are readily observable and measured with a high degree of accuracy. For example, Duflo (2004) estimates the effect of an Indonesian school construction program on educational attainment. Both the outcome of interest, school attainment, and the treatment, the number of new schools built in each district-year, are obtained from household surveys or public records. The mapping of variables of interest to data objects in this case, is straightforward.

In other cases, outcomes and individual characteristics are underlying traits not easily observed by the researcher. The importance of measuring outcomes is self-evident, but other characteristics may also play an important role since, as argued above, an important part of evaluation is the identification of the mechanisms that lead to programme impacts. In order to do so, the researcher needs to understand the way in which individuals behave and the drivers of such behaviour. For example, expectations about the potential returns to an investment will shape how agents respond to a policy aiming to boost that investment.

An early example of these efforts to measure underlying factors driving economic behaviour is the study of income expectations. As part of a telephone survey of households, Dominitz & Manski (1997) asked respondents what they thought the likelihood was of their income falling within certain bands. This novel measure provided researchers with not only the mean, but also the variation in expectations, which was correlated with other observable characteristics such as age, income and employment status. Probabilistic measures of this sort are extremely valuable for impact evaluations, since they can provide insights into the decision process of economic agents and inform the way we model their behaviour (Manski, 2004).

The identification of the mechanisms that determine the outcomes of interest is particularly difficult when some of the mediating variables are the results of individual choices and therefore affected by a variety of other factors, some of which can be correlated to the outcomes of interest. This 'endogeneity' problem is difficult to solve. However, ignoring the influence of these core issues on policy effectiveness is problematic and limits the insight of the evaluation exercise.

New measures can capture important aspects of behaviour and, therefore avoid arbitrary assumptions about individual behaviour. Parental investment, for instance, is likely to be driven in part by parental ideas about its usefulness. Therefore, an evaluation of the impacts of a policy on parental child investment would benefit from measuring these expectations. Instead of simply reporting the overall results of the policy, researchers could also test whether it changed parental beliefs, and further

understand the channels for policy effectiveness (or lack thereof). This could then lead to improvements in policy design that would otherwise not have been possible.[1]

Building novel measures to capture expectations, risk perceptions, beliefs, attitudes, or other unobserved phenomena (e.g. intra-household bargaining processes) can be extremely useful for evaluation.[2] But these measures come at a cost. First, they often require the collection of additional variables in the field, leading to longer, more expensive surveys. More importantly, they also require an understanding of the economic incentives and constraints under which agents operate, or an explicit model of human behaviour with which to model the relationships between the variables in question.

Additionally, these measures must be practical, should be readily implementable and, in cases when they will be embedded in household surveys, easily understood by respondents, to produce reliable results. The challenge is therefore to develop measures that help identify and estimate core parameters of a model that researchers want to estimate, and that fulfil the empirical requirements for the application of appropriate estimation methods. Such measures not only improve a researcher's estimates of impact, but also shed light on the mechanisms behind such impacts. In the remainder of this section, we describe examples in which new measures were developed in the context of impact evaluations, and the way in which they assisted our understanding of the underlying phenomena.

---

[1] An example of how to incorporate measures of parental expectations on a child's human capital accumulation over the life cycle can be found in Attanasio et al. (2017).

[2] These measures are often designed at the level of the agent, but other, higher-level variables such as district or village level observable characteristics may also prove helpful.

Example 1: measuring parental investment and beliefs

Parental behaviour and investment is a fundamental driver of child development. This fact is very much accepted, and it is particularly important in the early years of life (Heckman & Mosso, 2014). The stimuli children receive in a variety of dimensions affect their cognitive and language development, their executive functions as well as their socio-emotional skills. And yet, many parents do not seem to provide enough of such stimuli. An important research agenda in this field, therefore, is the characterisation of the drivers of parental investment. In order to do so, many measurement challenges are extremely salient. First of all, it is necessary to have reliable measures of 'parental investment'. By parental investment we mean both materials, such as toys and books, and time, such as the time spent by adults in stimulating and interacting with children.

As far as materials are concerned, it is important to collect information not only on commodities that parents might buy but also on alternative sources that might be used for play and other child activities. In this respect, there are now a number of measurement tools that make direct reference to a variety of materials. Based on the Home Observation for Measurement of the Environment (HOME) inventories designed by Bradley & Caldwell (1977), a host of different versions have been developed for children of different ages.

Reliable measures of time investment are much harder to collect. A particularly challenging and important dimension is the quality of the time spent with the child. A certain amount of time simply spent *with* the child but not engaging with her does not have the same effect of the time spent with child doing specific activities, such as playing, talking or reading books. In this respect, although useful, time use diaries might

have very limited information unless specific activities are considered. Recent efforts have developed a number of more nuanced measurement tools to collect information on specific activities and characteristics of child-parent interactions. One recent review of such measures is Dallay & Guedeney (2016).

In addition, more and more often, researchers use data that include direct observations of child-parent interactions from structured or unstructured sessions. Information on the degree of attachment, the quality of interactions, and the number of words used, among other measures, are all of interest when seeking to characterize child-parent interactions. To that end, recordings from these sessions are then analysed using a standardized coding system to arrive at meaningful and statistically useful measures. One example of these coding systems in the revised Family Observation System (FOS) proposed by Dadds & Sanders (2012).

Further innovation in new measures of the quality and quantity of parental investments could push this research agenda forward. We believe this is a worthwhile endeavour. More importantly, arriving at standardised measures to assess parental abilities and their interactions with the children, will be key to allow comparisons across studies and contexts, and to understand whether findings replicate over time or not.

An important driver of parental behaviour and investment is the parents' perception of the process of child development and of the usefulness of their investments. A small amount of literature indicates that parents from poor socio-economic backgrounds often consider investment not particularly useful and think that children develop naturally without any specific input from adults or their environment. In her book *Unequal Childhoods: Class, Race and Family Life,* Annette Lareau proposes alternative models that parents might have: the pursuit of natural growth used by poor and working class parents versus concerted cultivation used by middle class

families (Lareau, 2003). From a quantitative point of view, since such beliefs as are likely to inform parental behaviour, it is important to have the possibility of measuring them.

In recent work, Cunha et al. (2013) and Attanasio et al. (2009) developed new measurement tools to elicit quantitative measures of such beliefs. The authors present different investment scenarios (and current development of the child) and ask the parents of participating children to quantify the likely development of a hypothetical child under the different scenarios. These questionnaires give direct measures of the perceived rates of returns of certain investments under specific conditions and, under some assumptions, can be used to infer complete visions of the developmental process.

Obviously, such measures do not provide a complete picture of parental behaviour but are useful complements to standardised measures that include parental budgets and financial resources. In addition to such measures, it is also useful to collect information on possible stress factors that might reduce parental ability in their interactions with children.

Measures of parental beliefs can be used both in the characterisation of parental investment and as intermediate outcomes. The latter case is particularly useful in situations where researchers attempt to unpack the effect of an intervention. Suppose that, as a result of a particular intervention, researchers observe that parents increase their investment in their children. A legitimate question to ask is why they do so. One possible answer is that they change their perception of the usefulness of certain investments and interactions. Another possibility is that parents of children who were struggling in school decided to help them with their homework more often. In order to understand the medium and long term implications of the changes observed after said

intervention, it would be important for researchers to be able to identify which of the possible mechanisms is correct. Measures of parental beliefs are essential for this.

As in the case of parental investments, further research into the design of new, more accurate measures of parental beliefs is important to push this research agenda forward. Again, arriving at a set of agreed measures is also key, in order to allow for comparisons across studies and contexts. Researchers should also keep in mind that the meaning of these new measures may change over time and be sensitive to policy interventions, as the underlying factor they aim to capture evolves.

Example 2: measuring intra-household bargaining power

Most conditional cash transfer (CCT) programs around the world select a woman in the household to be the recipient of the transfer (Fiszbein & Schady, 2009). The argument frequently used in support of targeting transfers to women is not only that transfers promote gender equality and empower women, but that through the empowerment of women, they benefit children as well.

Theoretical models and empirical studies also reveal that such targeted transfers make a difference to household choices and outcomes indicating that the targeting does in fact empower women (Thomas, 1990; Hoddinott & Haddad, 1995; Lundberg, et al., 1997; Browning & Chiappori, 1998; Ward-Batts, 2008; Attanasio & Lechene, 2002, 2014; Doss, 2006). However, there is no clear consensus on the precise mechanism through which households make decisions and allocate consumption when receiving a cash transfer, and there is limited evidence on the exact mechanism linking money transfers targeted to women and empowerment within the household.

In the past few decades, many surveys have included batteries of questions aimed at measuring the extent to which women are empowered within the family. These measures are also used for some evaluations on the effect of empowerment programs. A

typical set of questions, used in many different contexts, asks respondents to identify who is in charge of certain decisions determining, for example, expenditures on different household consumption items, schooling, or various investments. The Demographic and Health Surveys (DHS), conducted in over 90 countries, is an example of a large study that often includes a module with such questions. Possible answers to these questions are that the wife is in charge, the husband is in charge, or spouses decide jointly. In many datasets, answers to these questions are bunched on the `both' categories, and very limited variation is obtained.

In the context of conditional cash transfers, for instance, the PROGRESA evaluation survey included several of these questions. This CCT did not seem to have shifted the answers to these questions (see for instance Adato et al. (2000)). Therefore, if one were to interpret those results literally, one would conclude that the transfer program, despite offering significant transfers to women, did not empower them. However, as empirical studies of consumption reveal that the targeted transfers induce households to make different consumption choices, a better measure of female influence in household is called for.

Almås et al., (2018) suggest a complementary measure of the relative bargaining strength or of women within the household. Rather than relying on traditional survey questions about who makes certain decisions regarding resource allocation within the household, they directly measure women's willingness to pay to gain control over income through an incentivised experiment. This experiment was implemented in urban areas of Macedonia. The women selected to participate were presented with a sequence of choices, between an amount Ak for themselves and an amount Bk for their husband (where Ak is usually smaller than Bk). The sequence of choices is designed to identify the value that makes the participants indifferent between receiving Ak and their husband

receiving *Bk*. A woman with a large degree of influence over household expenses is expected to choose quantity *Bk*, in order to maximize her (and her household's) income. The experiment therefore elicits the participant's willingness to pay to become the recipient of a cash transfer offered to the household. Said variable can be interpreted as a measure of power or influence within the household.

The researchers then implement this lab-based measure on a sample of women participating in a Macedonian conditional cash transfer (CCT) scheme that randomly assigned the transfer to either the woman or the man in each household. They find that the willingness to pay estimated from their lab experiment is smaller among women who received the transfer themselves, than among the women who did not. They thereby conclude that women who received the CCT have higher discretion or bargaining power within their households. Conducting similar evaluations of measures based on traditional survey questions about household decision-making, such measures point in the same direction, but the estimates are more imprecise and not significant for all survey modules. As such, this lab-based measure serves as an example that the laboratory may be an important additional tool worth exploring in settings where the measurement challenges seem hard to solve in the field setting alone.

New measures using publicly available datasets

Innovations in measurement have also been carried out in other fields where the challenge is not the absence but rather the overwhelming abundance of data. Examples of this are the recent contributions to the environmental economics literature. Over the last decade, remote sensing instruments in orbit have recorded a wide range of atmospheric variables at high spatial and temporal resolutions. This allowed for the development of almost real-time estimates of weather conditions, as well as atmospheric concentrations of key pollutants (NOx, SOx and PMx concentration, for

example). Researchers have found multiple ways of incorporating this information as inputs for their analysis, as both controls and outcomes (Donaldson & Storeygard, 2016).

To exploit the advantages provided by this glut of data, researchers must find ways of synthesising the information contained in these high dimensional datasets into statistically useful and economically meaningful objects. For example, they might only be interested in establishing a causal relationship that allows for the use of instrumental variables in a clean and tractable way. An example of this is the use of total precipitation as an instrument for economic shocks, as used by Miguel et al (2004). Other examples are the use of binary variables to reflect the exposure to a certain event, such as a drought, irrespective of the severity of each event.

While practical and parsimonious, these measures may be too coarse for some applications, leading to an under-rejection of null hypotheses (Hsiang, 2016). A more complete characterisation of socio economic responses to complex phenomena such as weather or other environmental shocks requires the construction of more detailed measures. The response to environmental phenomena might be non-linear in nature, and its characterisation might need to accommodate asymmetric effects, thresholds, and incremental or accumulated effects. Recent efforts in the literature have provided a wide range of examples of these measures. The use of high order polynomials of average temperatures, for example, and of completely flexible `binned' approaches are examples of this (Schlenker & Roberts, 2006). Several authors have noted the possible non-linear nature of environmental impacts on human activity, and thus the importance of including estimates of higher order moments of their distribution. In studying the impacts of precipitation on Indian crop yields, for example, Fishman (2016) showed the

importance of considering the intra-seasonal distribution, as well as the total amount of rain.

There have been significant advances in the development of environmental measurement variables over the last two decades, of which we have mentioned only a few examples above. For the purposes of identifying and contrasting policy impacts in different contexts, a researcher would ideally want to account for a wide range of environmental phenomena that may be affecting her results. This would be especially useful to explore all the possible channels through which policy might influence her outcomes of interest. However, as we have seen in this section, this sometimes clashes with the objective of creating simple, communicable results, or of establishing simple causal relationships. The development of compact, statistical objects that are both meaningful and easily implementable is a field of active research in economics. Researchers involved in impact evaluations may find large benefits from collaboration with other disciplines, such as agricultural and environmental sciences, as well as social sciences beyond economics, in the process of continuous refinement of these measures.

A final point worth considering is the importance of coherent measurement across contexts. While it might be beneficial for researchers to tailor their measures to the context of their interventions, it is also important, from the policy maker's point of view, to be able to compare results from a wide range of evaluations from different contexts. A trade off soon arises between the development of ideal measures for each particular context, and the possibility to conduct meaningful inference from multiple studies. Researchers working in the development of new measures must keep this trade off at the forefront of their work, in order to balance nuance with coherence.

Using multiple measures

In many applications of social science, the search for a single, ideal measure for a

particular object of study might prove to be futile. Preference is generally given to the use of few measures for the sake of interpretative clarity and model parsimony, but, often, this may come at a significant cost in terms of measurement error. This error might be classical, in the sense that the measure used contains large amounts of 'noise', attenuating impact estimates, or non-classical, if the measure used omits an important underlying trait that researchers aim to measure appropriately, introducing bias into their estimations. Both sources of error are problematic from the point of view of impact evaluations, and may lead researchers and policymakers to erroneous conclusions regarding policy effectiveness.

Instead of relying on a single measure, in some cases, it might be more appropriate to work with two (or more) imprecise measures. These might be measures of outcome or control variables. Rather than investing large amounts of resources in the design and implementation of a theoretically ideal measure, multiple, less expensive measures might be available that can allow researchers to achieve similar degrees of insight and certainty. These measures will be noisy by nature, but to the degree that they are not perfectly correlated between themselves, they may help capture the underlying concepts more accurately, and control for their respective sources of measurement error.

Measurement is expensive. Designing and collecting data takes up large amounts of time both from researchers and survey participants. At the same time, there is a wealth of publicly available, freely accessible datasets that can be combined to survey data purposefully collected to conduct impact evaluations. Broadly speaking, these online datasets can be grouped as follows: 1) repositories of household and individual surveys (e.g. DHS data); (2) gridded datasets (e.g. geographical incidence of conflict, natural disasters); and (3) regional/country level datasets (incidence of Intimate Partner Violence at country level, migration flows). The Annex to this paper provides

an inventory of several online resources grouped under each category and provides examples of recent studies that have leveraged some of these datasets as part of impact evaluations. In the remainder of this section, we focus on the problem of how to make sense of the information contained in multiple measures.

Before proceeding with this, an important distinction must be made. Consider the case of multiple outcome measures. If a researcher is interested in testing for the effect of a policy on multiple distinct outcomes, unrelated between each other, the standard procedure for robust causal inference is to conduct individual *t*-tests of mean comparison of outcomes. Adjustments to the estimated standard errors for these individual tests are necessary to control for the increased probability of false positive detection.[3] An example of this could be testing for the effects of a new drug on the prevalence of two, unrelated, health conditions.

This section addresses a related but different scenario: one in which a group of outcomes is believed by the researcher to represent multiple measures of a single underlying factor of interest. In these cases, composite indices are often constructed using the multiple outcome variables, and inference is conducted on the indices, instead of, or in addition to, the individual measures themselves. Most policy evaluation specialists will be familiar with composite indices of this kind. In the next section, we do a brief description of the main methods used to construct them, the assumptions behind them, and the ways in which they address measurement error.

Construction of indices from multiple measures

An example of a composite index that most researchers working in economic

---

[3] These adjustments are discussed more at length in Section 5.

development will be familiar with is the relative wealth asset index, proposed by Filmer & Pritchett (2001). Using 21 survey responses on home ownership of common durable assets, characteristics of the dwelling and land ownership, the authors construct an index that ranks households along wealth lines. The variables used are assumed to be noisy measures of an underlying factor, in this case, long-term household wealth. Filmer & Pritchett's index is built using the first principal component of a principal component analysis (PCA). It assumes that there is an underlying "signal' driving these responses, in this case, long-term wealth, and produces the linear combination of answers that reflects the highest possible share of this signal. The result is a wealth ranking of households that has no cardinal interpretation, but as the authors show, performs well at classifying households into wealth bins, when compared to other available measures such as consumption.

PCA identifies the linear combinations of responses that maximise the explained variation contained in them. The first principal component is the combination that explains the most variation; the second one is the one which explains most of the variation left, and so on. The components are orthogonal to each other, so each successive one is designed to explain as much of the *remaining* unexplained variation as possible.

Composite indices have also been used to construct measures of child development. Researchers have designed a battery of tests for children that provide measures for different skills at particular stages of a child's life cycle. These assessments are less prone to misreporting than household income, but none of these tests are free of measurement error in the sense that they, at best, provide relatively noisy signals of the child's cognitive or non-cognitive development.

A first principal component index as the one described provides two main advantages. First, it provides a more precise measure of the assumed underlying factor (in our first example, long-term wealth) than what can be inferred from the set of individual measures, by removing noise, or measurement error. Second, it provides a single measure with which to rank households or individuals which is easily interpretable and applied. The challenge for practitioners is to ensure that the indices constructed indeed reflect some underlying trait or characteristic. This is an important caveat to keep in mind before deciding to use composite indices of any type. Indices that do not target meaningful, theoretically based concepts are bound to lead to confusion and will fail to provide clarity. Researchers should ensure that the indices they construct are the appropriate empirical counterparts of theoretical parameters of interest from their underlying estimated model.

In most cases, this can be verified during the construction of the indices themselves. For example, in the construction of a long-term household wealth index, one would expect that durable consumer goods such as refrigerators or computers would enter positively, meaning that their ownership increases the position of a household in the ranking. The opposite is true for characteristics of variables that suggest low levels of income or access to infrastructure, such as dirt floors or no access to electricity. These components should enter the index in a negative way, i.e. reducing a household's position in the ranking. These checks are key to ensure that meaningful indices are being constructed.

PCA indices also have a series of limitations. First of all, they are not scale invariant. It is therefore important to standardize the values of the variables used to construct a PCA index, whenever they are not all expressed in the same scale. This excludes, for example, the case of PCA indices constructed using a set of binary

variables. Second, it is not obvious that a single component (i.e. the first one) will always capture the underlying trait that is assumed to be behind the set of variables being used to construct it. Filmer & Pritchett analyze carefully how the variables selected enter into their first component, and conclude that this is the case in their dataset. However, some applications may justify the use of more than one component, or the estimation of indices for different subsamples. For example, certain assets may indicate low levels of wealth in urban areas, but may instead be indicators of higher wealth in rural areas. A pooled index for both rural and urban areas is not be recommended in this case.

A third shortcoming is that PCA indices perform poorly when ordinal discrete variables (such as the level of education of the household head) are included (Kolenikov & Angeles, 2009). More recent efforts have proposed more flexible methods for constructing indices, such as the polychoric PCA, factor analysis (FA), and Item Response Theory (IRT), among others. In each case, the objective is to extract an underlying factor or signal, from a series of measures which all contain different degrees of error. The index chosen for each application will depend on the specifics of the problem, but most of these methods now allow for more flexible measures to be included, such as test scores, or types of household construction, and are generally available in most statistical software packages.

Using data differently

The advent of large administrative, transactional and satellite datasets commonly referred to as Big Data, and the spread of low-cost processing power, present some unique opportunities to researchers working on impact evaluations. Indeed, recent work has already explored the possibility of leveraging these trends by incorporating ML

algorithms into the policy analysis workflow.[4] In this section we discuss the possibility of incorporating some standard ML tasks into policy evaluation, carefully considering their potential contribution, as well as their cost for researchers.

Understanding the role that these new tools might play in future work requires first a characterisation of the different problems they tackle, and how this can complement the existing paradigm. During the past two decades, programme evaluation techniques have mostly focused on identifying and measuring causal relationships. This issue lies at the core of any impact evaluation, where researchers aim to understand the effect of a policy (W) on an outcome (Y). In order to do this, researchers make assumptions about the data generating process, and establish, with varying degrees of flexibility, how variables affect each other. In the standard ordinary least squares (OLS) framework, for example, it is assumed that the relationship between outcomes, treatment and other variables is linear. Robust statements regarding causality and the identification of channels of impact also require a theoretical model regarding the possible sources of selection bias, and a careful experimental design that allows researchers to carry out meaningful inference. In most policy evaluation contexts, a randomised allocation of treatment plus careful policing (including blinding) for sources of post-randomisation confounding or a natural experiment allow economists to proceed under the assumption of unconfoundedness, or in other words, the assumption that after controlling for observable characteristics, treatment assignment is good as random.

ML algorithms, on the other hand, are mostly tailored towards *prediction*. In this sense, the objective of most machine learning tools is not to estimate a causal

---

[4] For the purposes of this paper, ML will refer broadly to the design of algorithms that optimize certain quantitative tasks, and draw heavily from the fields of statistics and computer science.

relationship, or a parameter in an underlying model, but to predict a given value y, given a series of observable characteristics *X*. ML methods often do not rely on any structural assumptions on the data generating process and allow the relationships between *X* and *y* to be completely flexible. Their objective is not to understand the mechanisms that determine *y*, but to achieve the most accurate out-of-sample prediction of *y* possible. ML's focus on prediction comes at the cost of less meaningful estimates of the underlying parameters, from a causal inference perspective (Belloni et al., 2014).[5]

The application of ML algorithms and Big Data into the workflow of policy evaluation researchers therefore hinges on identifying tasks in which they can complement and enhance the parameter estimation problem. Below, we describe some examples of recent efforts aimed at exploring this complementary role for ML in the context of impact evaluations.

Missing data

In recent years, significant advances have been made to convert large, often high dimensional, data products, such as satellite measurements, into practical and informative statistical objects. This proves especially useful in overcoming the problem, pervasive in developing countries, of missing data. A series of studies have tackled the problem of missing poverty and wealth data in developing countries using satellite datasets, which, compared to household surveys, have the natural advantage of global coverage. The first generation of these studies used publicly available 'nightlight' datasets, which contain images of night-time artificial lighting, and used them as

[5] The different nature of these two statistical tasks was first described by Leo Breiman, in his comparison of the data modelling culture, associated with econometrics, and the algorithmic modelling culture, more prevalent in data science (Breiman, 2001).

proxies for output (Chen & Nordhaus, 2011) and economic growth (Henderson, Storeygard, & Weil, 2012) at the country level. More recent efforts have combined satellite sensed datasets with ML algorithms to predict missing outcomes at a higher level of spatial resolution. Using a sample of five African countries covered by the World Bank's Living Standards Measurement Study (LSMS), Jean et al (2016) trained an algorithm that predicts cluster level consumption, expenditure and asset wealth from the LSMS, using publicly available satellite imagery as predictors. Notably, the authors find that their algorithm predicts of up to 75% of the variation across clusters. Importantly, they achieve superior predictive power for asset wealth than Blumenstock et al. (2015), who use individual mobile phone usage as a predictor, while only relying on publicly available datasets.[6]

These studies highlight the potential of new, publicly available data products and novel ML algorithms to overcome the problem of missing data when no reliable household surveys or censuses are available. These procedures allow researchers to map or predict characteristics at the local level, which may play important roles in mediating policy effectiveness, and may inform the decision of where to target future interventions. At the same time, these approaches have some important drawbacks related to their out-of-sample properties. The relationship between satellite data and household characteristics might vary by country or region, and over time. Researchers should not assume that the findings from one context translate to another, for which no survey data is available. This limits the range of applications for which they may be

[6] Arribas et al (2017) use similar input to predict an indicator of living environment deprivation for the city of Liverpool

useful to researchers working in policy evaluations, until further work refines the algorithms and incorporates other sources of data.

A particular source of measurement concerns arises when respondents have reasons (e.g. legal sanctions, fear of retaliation) to conceal or misrepresent their true responses to sensitive questions like evading taxes, taking bribes, or their degree of influence over household decisions. List randomisation (used for example in McKenzie & Siegel (2013) to measure illegal immigration), endorsement experiments (as used in Lyall, et al (2013) to understand support for the Taliban and foreign forces in Afghanistan), and reticence adjusted measures of corruption (as used in Kraay & Murrell (2013) to measure corruption in an enterprise survey in Peru), are possible alternate ways to elicit sensitive information of this kind. In other cases, courtesy biases or Hawthorne effects might result in data collection errors that are harder to avoid with conventional methods and can hardly be avoided in traditional interview settings. As noted by Peterson Zwane et al. (2011), these effects can be significant, highlighting the benefits from indirect and un-intrusive data collection methods. The advent of large transactional datasets which record regular transactions with no data collection needed opens important opportunities in this direction. Further research is encouraged in this direction to help elicit sensitive information, possibly combining multiple sources of data, and to avoid repeatedly approaching households in a survey context, to reduce the risk of making certain subjects more salient to respondents, and thus alter their behaviour.

Model selection

A routine task faced by researchers during an impact evaluation is that of selecting the appropriate set of control variables, or covariates, to include in the estimation of treatment effects. In non-experimental contexts, covariates are included in order to

comply with the unconfoundedness assumption, to assert that a certain treatment is as good as random after controlling for observable characteristics. In this case, the introduction of covariates aims at tackling omitted variable bias (OVB) and allows for a causal interpretation of the estimated treatment effects. In cases where a policy or treatment was randomly assigned, researchers also tend to include a series of covariates, besides treatment assignment, as controls. Their aim in this case is to increase the accuracy of their treatment estimates. Introducing additional covariates that are correlated with the outcome of interest reduces the overall residual variance, and often results in smaller confidence intervals on the parameter estimates (Duflo, et al., 2008). For example, introducing baseline values of the outcome of interest into the regression generally reduces the standard errors of the estimated treatment effects.

However, each additional covariate included in a regression model reduces its degrees of freedom and may end up inflating standard errors as a consequence. This is the case for covariates that provide little additional explanatory power, and therefore do not significantly reduce the variance of the regression residual term. Which covariates to include in a regression is thus not straightforward choice, and may significantly affect the conclusions of an impact evaluation.

The standard covariate selection approach is, in principle, based on theoretical considerations of the possible relationships between covariates and the outcome of interest. In practice, a set of individual or household level demographic controls are often included based on common practice and assumptions about possible mediating factors affecting policy effectiveness. Given the large amounts of covariates that may be available to researchers via surveys or secondary data sources, this approach may not be optimal, and improvements in the accuracy of the estimated treatment effects may be possible.

An ML-based alternative for the principled selection of covariates to include in the estimation of treatment effects is the use of the least absolute shrinkage and selection operator (LASSO). The standard LASSO approach aims at identifying the subset of regressors that best predict an outcome, including a penalty for each additional regressor included. In its standard formulation, LASSO coefficients are estimated by minimising the following expression:

$$\hat{\beta}_{LASSO} = \arg\min_{b} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\, b_j \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \qquad (1)$$

Where the first the first term is equal to the sum of the squared errors (SSE) of the ordinarily least squares (OLS) approach, and the second shows the penalty λ for each additional regressor with a non-zero coefficient.[7]

In the context of causal inference, LASSO may be a useful tool for selecting the best subset of control variables to include in a regression. Belloni, et al. (2014) propose a double selection procedure that uses a LASSO estimation to identify the best subset of regressors to predict (i) the outcome variable, and (ii) treatment assignment. Then they estimate treatment effects in a standard way, including the union of both sets of selected regressors, together with the treatment indicator. Using examples from past studies, the authors then show how their approach may lead to qualitatively different conclusions about the impacts of different policy changes, than those arrived by original authors

[7] As expected, results from LASSO covariate selection methods vary according to the value of the penalty parameter λ. Several methods have been proposed for the selection of an optimal λ, such as cross validation (James, et al., 2013).

selecting controls by intuition only.[8] Choosing the correct set of covariates may significantly increase statistical power, and allow researchers to more precisely estimate programme impacts.

The potential for such model selection methods to contribute to fruitful policy evaluation work increases with the number of covariates available to the researcher. The data-driven nature of the procedure also makes it replicable and reduces the margin for *cherry picking*.

A similar LASSO approach may prove useful in a different but related context, that of the selection of the optimal set of instrumental variables for some endogenous covariate. The intuition for this lies in the fact that the first stage in any two-stage least squares estimation is essentially a problem of prediction. Take, for example, the problem of measuring the effect of parental investments on child development. These investments will be correlated with a host of other determinants of child development, and a researcher might want to use an instrumental variable strategy to identify the causal effect of higher parental investments, keeping all other factors constant. A reasonable set of instruments to use would be local prices for child-related consumer goods, but these may be extremely large in number. Prices of, say, 50 different products, with all their possible interactions and lagged terms, can result in a set of covariates in the thousands. How should we select the set of prices that best predict parental investments?

---

[8] In another recent study, Bloniarz, et.al (2016) propose a different LASSO-based selection method and show that, even in an experimental context, where random assignment to treatment and control groups was largely successful, increases in the accuracy of ATE estimates is achieved. Farrel (2015) builds on the work by Belloni et al (2014) and proposes 'doubly-robust' estimator that allows for multi-valued treatments and imposes weaker assumptions on the underlying data generating process.

A LASSO approach, as suggested in (Belloni, et al., 2012) provides a principled way of selecting for the optimal number of instruments to include when presented with a large choice set and allows for conventional inference on the estimator of the instrumented variable. The second stage of the estimation can be carried out with standard estimation techniques, since in this case, LASSO variable selection is only applied in the first stage regression.

Finally, ML algorithms have also been proposed that conduct multiple estimations using alternative models and report a measure of the variation of the estimates across specifications. For example, Athey & Imbens (2015) propose principled and replicable ways of carrying out these kinds of tests and obtain a credible estimate of the sensitivity of program impacts to the specification used.

Heterogeneous treatment effects

Following the estimation of average treatment effects, researchers often ask who did the treatment work for, and who didn't it work for. From an academic point of view, identifying groups with different responses to a policy may help shed light on the possible channels of impact, and may help validate or reject existing theories about how and when that policy works. From the point of view of a policy maker, this is important for a more effective delivery, and for the efficient allocation of scarce resources.

The increased availability of administrative and transactional data and the reduced costs of data collection in experimental settings imply that researchers have access to a large number of observable characteristics for each unit of observation. While this has its obvious advantages, the study of heterogeneous treatment effects is hampered by the likelihood of detecting spurious, non-replicable, results. Traditionally, this has been addressed by establishing a pre-analysis plan (PAP), *ex ante* limiting the

number of dimensions to be explored, and by statistical corrections to significance tests for multiple hypothesis testing.

While ensuring internal validity, this approach has a series of limitations. First, in a standard PAP, the pattern that interactions between treatment assignment and other observable characteristics may take is limited. Rarely, if ever, do researchers contemplate the possibility of interactions between treatment assignment and covariates (or between the covariates themselves) taking any form other than a standard linear relationship. Second, and perhaps more importantly, researchers are (purposefully) constrained to study only the dimensions considered to be relevant for programme impacts before the intervention was carried out, and established in the PAP. Therefore, researchers rule out the possibility of exploring unexpected, but significant, patterns of interaction between other, not pre-specified, observable traits and programme impacts. In other words, the current approach to trial registration (intentionally) restricts how much researchers can learn from each impact evaluation.

In recent work by Athey & Imbens (2016), the estimation of heterogeneous treatment effects is presented as a prediction problem, and therefore one in which ML algorithms may play a valuable role. The innovation here is that instead of predicting an outcome (which can be readily observed), the algorithm estimates a conditional average treatment effect (CATE) for different sub-groups of the population. Building on early work by Breiman (2001), the authors propose a method that estimates the CATE for different subgroups in a recursive manner and allows for highly flexible relationships between variables. The procedure has tractable asymptotic properties, allows for the construction of confidence intervals for the estimated treatment effects, and is robust to concerns about false positives (Athey & Wager, 2017).

Causal forests, the ML method proposed by Athey and Imbens, have already been used for the estimation of heterogeneous treatment effects in both experimental (Davis & Heller, 2017) and non-experimental contexts (Saavedra & Romero, 2017). However, this method is still hard to implement using fixed effects regressions, clustered standard errors or continuous treatment variables, which are common in the policy evaluation workflow. Further development of the algorithm may soon incorporate these characteristics, making it more likely to be incorporated into the impact evaluation toolbox.

Additional examples

ML methods can be be of use in multiple stages of a policy evaluation. Consider for example, the question of designing survey instruments for a randomised controlled trial, subject to a budget constraint. Researchers aim to maximise the precision of the estimated treatment effect, and they can do this by increasing the number of interviews carried out, and by increasing the number of covariates collected. In the past, attention has focused on calculating an optimal number of interviews (McConnell & Vera-Hernandez, 2015) and survey waves (McKenzie, 2012) according to the nature of the outcome of interest. A recent study suggests an ML approach to tackle a related problem: the choice of the optimal number of questions to include in a survey. The authors address the trade-off between the marginal cost, in terms of interview time, and marginal benefit, in terms of additional explanatory power, of including each possible covariate in the survey, and define an ML algorithm that finds the optimal solution to that problem (Carneiro, et al., 2017).

Ludwig et al. (2017) propose an ML method for another common task in policy evaluation: testing the impact of an intervention on multiple outcomes. This applies to cases in which the intervention may have effects on several measures, such as health, or

on a single underlying factor, such as cognitive development, but that is measured using scores from several tests. Conducting multiple individual t-tests is problematic, because the likelihood of finding false positives (type I error) increases with the number of tests being carried out. Therefore, researchers adopt restrictive pre-analysis plans in which they set the outcomes they will test in advance, and correct the results of individual test statistics using methods now standard in the impact evaluation toolkit.[9] Ludwig et al propose a test that also controls for false positives, but that is more flexible than currently applied methods, by converting it into a problem of predicting treatment status based on the set of covariates, and thus exploiting ML's capacity for prediction.

The two approaches described in this section suggest additional directions in which ML methods may complement policy evaluation work. However, further work is needed to implement and to test their performance in practice, before incorporating them into the policy evaluation toolkit.

Discussion

There are tasks throughout the life cycle of an evaluation that may be understood as prediction problems, and for which ML algorithms may be useful in producing more credible and robust results. In this section we covered three examples, the production of missing data using remote sensing instruments, model selection using LASSO, and the study of heterogeneous treatment effects of a programme. While most of these examples are still in the proof-of-concept stage, and may require further work to become readily implementable tools, some, as in the case of LASSO selection of covariates, can now be easily applied using standard statistical packages. More importantly, as Belloni, et al.

[9] A reliable and not excessively conservative method for correcting standard errors for multiple hypothesis testing can be found in Romano & Wolf (2005).

(2014) demonstrate, the application of ML methods may lead to qualitatively different conclusions about policy effectiveness from those derived from standard theory driven model selection methods.

However, an important note should be made about the role that ML methods play in policy evaluation. As mentioned in a recent survey of the impact of ML on economics research, these new methods do not alter the important insights derived from the program evaluation literature regarding identification and causality (Athey, 2018). And in the same way that ML won't fix the fundamental problem of causal inference, that is, how to assign and measure causal impacts, it won't change mediation analysis either. Observing higher programme impacts among a sub-sample of agents with a particular set of characteristics will still be just that, the observation that programme effectiveness is correlated with a set of covariates, and nothing more. Researchers are thus encouraged to explore the potential of ML methods to improve the accuracy of programme impact estimates, keeping in mind the insights of the impact evaluation toolkit.

Conclusion

Data and measurements are key inputs into the design of effective, poverty alleviation policies. In this paper we discussed three categories of related approaches to innovation in data and measurements relevant to policy evaluation practice, provided a few detailed examples of their application, and drew some recommendations for researchers.

We began by discussing some recent advances in the construction of novel measures. First, we described the case of parental investments in child development, and the challenges involved in measuring both the time and quality of the time they spent together, as well as parental beliefs and expectations. The second example came from the intra-household bargaining literature and describes recent efforts at measuring

more precisely the level of empowerment wives have within their household. These examples shed light on the limitations of existing measures and suggest possible avenues of research for further improvement in these and other areas. Future efforts in the development of new measures should proceed with an awareness of the trade-off between nuance and coherence across studies. New measures that address the core parameters of underlying theoretical models have the potential to significantly increase the insights derived from an evaluation. At the same time, researchers should keep in mind that the ability to compare results across different contexts is extremely important for policymakers. Data harmonization across studies is challenging, if at all possible, so new measures should be designed with the objective of being applicable to a wide range of contexts.

The second category of approaches discussed relates to the use of multiple measures. When ideal measures are hard to come by, or a trait does not lend itself easily for measurement in a single dimension, several measures might provide better results. We described how the construction of combined indices, using multiple measures can be used, the advantages and limitations of this approach. Researchers must pay special attention to how the components of the indices constructed enter into the indices themselves, to ensure that they indeed represent measures of the core parameters of the models they are interested in.

Finally, we presented examples of tasks for which ML algorithms may complement a researcher's work in policy evaluation. These include model selection, the selection of instrumental variables and the study of heterogeneous treatment effects. The potential of these algorithms lies in the fact that they can provide a principled way to address issues of optimal design or selection when economic theory alone cannot provide guidance. While ML methods do not alter the key insights of the programme

evaluation literature regarding causal inference, incorporating ML methods into the policy evaluation workflow can result in more robust and credible estimates, and increase the transparency of our results.

References

Adato, M., De la Briere, B., Mindek, D., & Quisumbing, A. (2000). *The impact of PROGRESA on women's status and intrahousehold relations.* Washington D.C.: International Food Policy Research Institute.

Almås, I., A. A., Attanasio, O., & Carneiro, P. (2018). Measuring and Changing Control: Women's Empowerment and Targeted Transfers. *Economic Journal*.

Arribas-Del, D., Patino, J., & Duque, J. (2017). Remote Sensing-based Measurement of Living Environment Deprivation: Improving Classical Approaches with Machine Learning. *PLoS ONE*.

Athey, S. (2018). The Impact of Machine Learning on Economics. In *Economics of Artificial Intelligence.* Chicago: University of Chicago Press.

Athey, S., & Imbens, G. (2015). A Measure of Robustness to Misspecification. *American Economic Review*, *105*(5), 476-480.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 7353-7360.

Athey, S., & Wager, S. (2017). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*.

Attanasio, O., & Kaufmann, K. M. (2009). Educational choices, subjective expectations and credit constraints. *NBER Working Paper No 15087*.

Attanasio, O., & Lechene, V. (2002). Tests of income pooling in household decisions. *Review of economic dynamics*.

Attanasio, O., & Lechene, V. (2014). Efficient responses to targeted cash transfers. *Journal of Political Economy*.

Attanasio, O., Meghir, C., Nix, E., & Salvati, F. (2017). Human capital growth and poverty: Evidence from Ethiopia and Peru. *Review of Economic Dynamics,* 234-259.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012, November). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica, 80*(6), 2369-2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives, 28*(2), 29-50.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., & Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 7383-7390.

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

Bradley, R. H., & Caldwell, B. (1977). Home observation for measurement of the environment: a validation study of screening efficiency. *American Journal of Mental Defficiency*, 417-420.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199-231.

Browning, M., & Chiappori, P. (1998). Efficient intra-household allocations: A general characterisation and empirical tests. *Econometrica*.

Carneiro, P., Lee, S., & Wilhelm, D. (2017). Optimal Data Collection for Randomised Control Trials. *cemmap Working Paper, CWP45/17*.

Chen, X., & Nordhaus, W. D. (2011, May). Using Luminosity Data as a Proxy fro Economic Statistics. *Proceedings of the National Academy of the Sciences, 108*(21), 8589-94.

Cunha, F., Elo, I., & Culhane, J. (2013). Eliciting Maternal Expectations about the Technology of Cognitive Skill . *NBER Working Paper No 19144*.

Dadds, M. R., & Sanders, M. R. (2012). *Behavioural Observation Coding System: FOS-V.* Queensland, Australia: Griffith University.

Dallay, E. G., & Guedeney, A. (2016). Parent-Infant Interaction Assessment. In A.-L. Sutter-Dallay, N.-C. Glangeaud-Freudenthal, A. Guedeney, & A. Riecher-Rossler, *Joint Care of Parents and Infants in Perinatal Psychiatry* (pp. 93-108). Springer, Cham.

Davis, J. M., & Heller, S. (2017). Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs. *American Economic Review: Papers & Proceedings*, 546-550.

Dominitz, J., & Manski, C. F. (1997). Using Expectations Data to Study Subjective Income Expectations. *Journal of the American Statistical Association, 92*(439), 855-867.

Donaldson, D., & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives, 30*(4), 171-98.

Doss, C. (2006). The effects of intrahousehold property ownership on expenditure patterns in Ghana. *Journal of African economies*, 149-180.

Dreibelbis, R., Freeman, M. C., Greene, L. E., Saboori, S., & and Rheingans, R. (2014). The impact of school water, sanitation and hygiene interventions on the health of younger siblings of pupils: a cluster randomized trial in Kenya. *American Journal of Public Health*, 91-97.

Duflo, E. (2004). The Medium Run Effects of Educational Expansion: Evidence from a Large School Construction Program in Indonesia. *Journal of Development Economics, 74*(1), 163-197.

Duflo, E., Glennerster, R., & Kremer, M. (2008). Using Randomization in Development Economics Research: A Toolkit. In T. Schultz, & J. Strauss, *Handbook of*

Development Economics Vol 4 (Vol. 4). Amsterdam and New York: North Holland.

Farrel, M. (2015). Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations. *Journal of Econometrics*, 1-23.

Filmer, D., & Pritchett, L. (2001). Estimating wealth effects without expenditure data-or tears: an application to educational enrollments in states of India. *Demography*, 115-132.

Fishman, R. (2016). More uneven distributions overturn benefits of higher precipitation for crop yields. *Environmental Research Letters, 11*, 1-7.

Fiszbein, A., & Schady, N. R. (2009). *Conditional cash transfers: reducing present and future poverty*. Washington D.C.: World Bank Publications.

Heckman, J. J., & Mosso, S. (2014). The Economics of Human Development and Social Mobility. *Annual Review of Economics*, 689-733.

Henderson, J., Storeygard, A., & Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review, 102*(2), 994-1028. doi:10.1257/aer.102.2.994

Hoddinott, J., & Haddad, L. (1995). Does female income share influence household expenditures? Evidence from Côte d'Ivoire. *Oxford Bulletin of Economics and Statistics*, 77-96.

Hsiang, S. (2016). Climate econometrics. *Annual Review of Resource Economics*, 43-75.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New Yok: Springer.

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

Kirchberger, M. (2017). Natural disasters and labor markets. *Journal of Development Economics*, 40-45.

Kolenikov, S., & Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth*, 128-165.

Kraay, A., & Murrell, P. (2013). Misunderestimating corruption. *World Bank Policy Research Working Paper Series*.

Lareau, A. (2003). *Unequal childhoods: class, race and family life.* University of California Press.

Ludwig, J., Mullainathan, S., & Spiess, J. (2017). Machine learning tests for effects on multiple outcomes. *Working Paper*.

Lundberg, S. J., Pollack, R. A., & Wales, T. J. (1997). Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *The Journal of Human Resources*, 463-480.

Lyall, J., Blair, G., & Imai, K. (2013). Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan. *American Political Science Review*, 679-705.

Manski, C. F. (2004). Measuring expectations. *Econometrica, 72*(5), 1329-1376.

McConnell, B., & Vera-Hernandez, M. (2015). Going beyond simple sample size calculations: a practitioner's guide. *IFS Working Papers, W15/17*.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics, 99*(2), 210-221.

McKenzie, D., & Siegel, M. (2013). Eliciting illegal migration rates through list randomisation. *Migration Studies*, 276-291.

Miguel, E., Satyanath, S., & Sergenti, E. (2004). Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy, 112*(4), 725-753.

Peterson Zwane, A., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., . . . Banerjee, A. (2011). Being surveyed can change later behaviour and related parameter estimates. *Proceedings of the National Academy of Sciences, 108*(5), 1821-1826.

Romano, J., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 1237-1282.

Saavedra, S., & Romero, M. (2017). Local Incentives and National Tax Evasion: The Response of Illegal Mining to Tax Reform in Colombia.

Schlenker, W., & Roberts, M. J. (2006). Nonlinear Effects of Weather on Corn Yields. *Review of Agricultural Economics, 28*(3), 391-398. doi:10.1111/j.1467-9353.2006.00304.x

Thomas, D. (1990). Intra-household resource allocation: An inferential approach. *Journal of human resources*.

Ward-Batts, J. (2008). Out of the wallet and into de purse: using micro data to test income pooling. *The Journal of Human Resources*, 328-351.

Appendix - A survey of useful public datasets

This annex provides a guide to online data sources which are either (1) repositories of household and individual surveys; (2) gridded datasets; or (3) regional/country level datasets. The list is by no means exhaustive, as the availability of public datasets is continuously increasing. The purpose of this Annex is only to serve as a starting point for researchers working in impact evaluations who might be considering the possibility of using public data to evaluate interventions, or to complement their survey data.

Most of these datasets include geographical coordinates that enable them to be linked to users' own data, and represent powerful resources to enrich evaluation studies through time-varying and spatially varying layers of information that would otherwise be costly to collect. For example, they provide information on the socio-political setting, trends in human events such as migration flows and incidences of natural phenomena.

Linking information from state administrative databases with survey data on individuals/households also has the potential to improve the measurement of program participation and of program services actually received. In the empirical analysis of the evaluation, variables that are derived from these sources can be used as instruments, provide rigorous identification strategies, or help identify natural assignment rules to treatment. This approach may be limited though, by the availability of administrative datasets, and by the large share of unregistered work that is common in many developing country contexts.

A few examples from peer reviewed studies illustrate the use of these resources. Dreibelbis, et al (2014) examines the impact of school water, sanitation, and hygiene (WASH) interventions on diarrhoea-related outcomes among younger siblings of school-going children. The authors use data from the 2003 Kenya Demographic and Health Survey (DHS) to calculate baseline prevalence rates of diarrhoea among children

younger than 5 years, during the last week. Using these rates, calculated at the cluster level, the authors then randomly assigned the intervention among clusters.

Another example comes from Kirchberger (2017), which explores how a large earthquake in Indonesia affected labour market outcomes (evolution of wages across sectors). To do so, the author combines data from several sources. To measure exposure to the earthquake, the paper uses data on reported damages from earthquakes sourced from the *DesInventar* database for Indonesia, maintained by the *Indonesian National Board for Disaster Management*. The database contains information on the type of disaster and geographical location, as well as its effect on human life and damage to property and infrastructure. To measure the strength of the disaster, the paper employs the number of destroyed houses and district-level population data from the 2005 Indonesia's Statistical Agency to account for differences in population density across districts.

It is important, however, to bear in mind some of the limitations inherent to these data. Administrative data, for example, can be of poor quality because data-tracking systems are often decentralised with few quality-control mechanisms in place. These data are not necessarily collected for research purposes, and do not always contain a sufficient level of detail. A second limitation is around the representativeness and reliability of some of these data. For example, the accuracy and reliability of temperature and precipitation measurements (which are often used as instrumental variables) may be compromised during episodes of civil conflict. Keeping these caveats in mind, recent efforts show the large potential that administrative and other publicly available datasets have, to provide additional measurements of relevant variables, when evaluating programme impacts.

Household and Individual Surveys

- The Poverty and Action Lab (J-Pal) lists over 800 RCT studies published by J-Pal as of October 2017. Of these, 175 include downloadable datasets which are publicly available: https://www.povertyactionlab.org/evaluations?f[0]=field_external_data:title:1
- The World Bank has an Impact Evaluation Microdata Catalogue which gives access to 113 datasets (71 of which are public use data files) and metadata underlying IEs conducted by the Bank or partner agencies: http://microdata.worldbank.org/index.php/catalog/impact_evaluation
- The International Initiative for Impact Evaluation makes available 31 datasets used in completed IE studies. https://dataverse.harvard.edu/dataverse/3ie
- The International Food Policy Research Institute (IFPRI) offers open access to 11 country-level data sets that were used to conduct baseline surveys (1999 - 2016): http://www.ifpri.org/collections/related/publication_tools/26
- The Programme Quality Team of Oxfam GB has made available for download 16 survey datasets which were used to carry out their Effectiveness Reviews: https://views-voices.oxfam.org.uk/methodology/real-geek/2016/09/real-geek-out-in-the-open-oxfams-impact-evaluation-survey-data-now-available-for-download
- The Low and Middle Income Longitudinal Population Study Directory (LMIC LPS Directory), developed by the Institute for Fiscal Studies (IFS), is a searchable directory of 173 longitudinal studies from low and middle-income countries with a sample size of 500 households or more: https://www.ifs.org.uk/tools_and_resources/longitudinal

Gridded Datasets

Humanitarian/Conflict

- Armed Conflict Location and Event Data Project (ACLED) is a geo-spatial database on conflict that tracks the actions of opposition groups, governments, and militias across Africa and Asia, specifying the exact location and date of battle events, transfers of military control, headquarter establishment, civilian violence, and rioting. ACLED data are disaggregated by type of violence including battles between armed actors, violence against civilians, and rioting and a wide variety of factors including government forces, rebel groups, militias, and civilians. Available at: https://www.acleddata.com. The following link collects many publications where the data have been used: https://www.acleddata.com/research-and-publications/. See also https://www.acleddata.com/wp-content/uploads/2015/10/Conflict-Datasets-Typology-Overview-Regional1.pdf for additional conflict and violence dataset publicly available
- This database is managed by the Peace Research Institute Oslo (PRIO) and includes both replication data and original datasets: https://www.prio.org/Data

Climate

- The Tropical Rainfall Measuring Mission (TRMM) is the result of the collaboration between NASA and the Japan Aerospace Agency, produces rainfall data at an almost global scale. Processed, clean datasets (Level 3) are available at a 0.25° spatial resolution and 3-hour time intervals, from 1998 to March 2015. Of course, with rainfall data, some work is needed to translate a quantity of rainfall into the experience of a shock, and that won't be as accurate

in some parts of the world as in others. Available at: https://pmm.nasa.gov/data-access/downloads/trmm

- The Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Terra satellite collects a wide range of data in its sun-synchronous polar orbit. Processed and clean (Level 3) datasets are available for land surface temperature, snow cover, aerosol/ozone concentrations, vegetation indices, forest fires and burned areas. Spatial resolutions vary by product but are generally as high as 0.05° and most products are available in daily formats. Available at: https://modis.gsfc.nasa.gov/
- The European Centre for Medium-Range Weather Forecasts (ECMWF) produces a range of climate reanalysis products, which combine data from weather monitoring stations with satellite and other sources and feeds them into a global climate model. Predicted values for most atmospheric and land surface variables of interest are available at high temporal resolution, in a 0.125° grid, from 1970 to today at: https://www.ecmwf.int/en/research/climate-reanalysis/browse-reanalysis-datasets
- NASA has developed their own climate reanalysis product, the Modern Era Retrospective-Analysis for Research and Applications (MERRA). Recently it has been replaced by its successor, the MERRA-2 product. It includes predictions for most atmospheric and land surface variables of interest in hourly, daily, and monthly formats, and a resolution of 0.5° x 0.625°. Available at: https://gmao.gsfc.nasa.gov/research/merra/
- The Palmer Drought Severity Index (PDSI) is a useful tool that assimilates multiple sources of weather data into a single index of exposure to drought risk.

Available at: https://climatedataguide.ucar.edu/climate-data/palmer-drought-severity-index-pdsi

- Global (land) precipitation and temperature from weather monitoring stations only, created by Matsuura and Wilmott, University of Delaware (UDEL). Accurate readings are available in a global 0.5° grid resolution, for areas with high density of monitoring stations. Where monitoring stations are rare or far apart, remote sensing datasets might provide more reliable readings. Available at: http://climate.geog.udel.edu/~climate/html_pages/download.html#T2014
- https://climatedataguide.ucar.edu/climate-data/precipitation-data-sets-overview-comparison-table is a repository of different datasets for precipitation data. It includes estimates of precipitation's distribution, amounts and intensity. This is important because precipitation can be fractal in space and discontinuous in time, unlike temperature which has a high degree of spatial and temporal correlation. Further, regional variations in topography can affect precipitation amounts significantly. Most precipitation data sets may be categorised into one of three broad categories: gauge data sets, satellite-only data sets, and merged satellite-gauge products.
- http://www.emdat.be Launched in 1988, the Emergency Events Database (EM-DAT) was created by the Centre for Research on the Epidemiology of Disasters (CRED). The EM-DAT database contains essential core data on the occurrence and effects of over 22,000 mass disasters in the world from 1900 to the present day. The database is compiled from various sources, including UN agencies, non-governmental organisations, insurance companies, research institutes and press agencies.

- http://www.desinventar.org/en/database The Disaster Inventory System - *DesInventar* is a conceptual and methodological tool for the construction of databases of loss, damage, or effects caused by emergencies or disasters. It includes database with flexible structure.

Agriculture

- IFPRI holds a repository of 12 datasets that allow the mapping of different agro-related identifiers onto survey data: http://ebrary.ifpri.org/cdm/search/collection/p15738coll3/searchterm/Geospatial%20Data/field/series/mode/all/conn/and/order/date
- The Stanford Centre on Global Poverty and Development (granular detail: geo-spatial level) (http://globalpoverty.stanford.edu/research/initiatives/data-development) has recently launched a "Data for Development" initiative, working to put in place strategies to combine different source of data to track the incidence of poverty-related phenomena, even in the remotest locations of developing countries (e.g. mapping satellite imagery over call data records to track disease outbreaks in remote villages, or the use of social media data to track the genesis of political protests and relate them to governments' responses to them).

Regional/Country-level Datasets

Gender

This remains a 'work in progress' area in terms of indicators available. Most publicly available data sources are at the regional level. DHS surveys (https://dhsprogram.com/Data/) are those typically mapped over survey data as they are representative at a regional level.

- Gender Based Violence (GBV) (granular detail: country level): https://unstats.un.org/unsd/gender/vaw A UN web portal provides data and detailed metadata for the two indicators for each country (proportion of women subjected to physical and/or sexual violence by a current or former intimate partner (IPV) in the last 12 months, and proportion of women subjected to sexual violence by persons other than an intimate partner since aged 15. Both datasets are available from May 2016 and can contribute to the monitoring of progress towards the achievement of SDG target 5.2.
- Gender and Development Basin Maps: http://maps.vista-info.net/gis/htm/IWMIBasinMaps/. These are gender-disaggregated data on measures of population, mortality, malnutrition and sanitation for the Nile, Volta, Ganges, and Mekong river basins. The maps provide accurate data on water and agriculture aspects, where gender inequities are known to be prevalent.

Migration

- The United Nations Global Migration Database (UNGMD), a comprehensive collection of empirical data on the number ("stock") of international migrants by country of birth and citizenship, sex and age as enumerated by population censuses, population registers, nationally representative surveys and other official statistical sources from more than 200 countries and territories in the world. Available at: http://www.un.org/en/development/desa/population/migration/data/index.shtml