Elements of quantitative research

Francisco Villamil

Research Design for Social Sciences
MA Computational Social Science, UC3M
Fall 2025

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Example: Wartime civilian deaths

Paper discussion

What is a research question?

- Any question we can answer
- Sometimes we say that we derive an RQ from a topic, and a theory from an RQ
  - → **Topic** > **RQ** > **Theory**
- In reality, an RQ can be thought of as an operationalization of an argument
  - → **Previous evidence** > **Argument** > **RQ** > **Hypotheses**
- **Even though** this ‘argument’ can be something anecdotical that we later develop into a proper, abstract theoretical argument
  - → And it would actually look into something like this:
  - → Previous > ‘Anecdotal argument’ > RQ > (Proper) Theory > Hs

What is a research question? Example

- Topic/idea: Charlie Kirk and the rise of political violence?
- 'Anecdotal argument'/intuition: online radicalization and young men more likely to commit individual acts of violence
- 'Main' RQ: Does online media cause an increase in lone-wolf political violence?
  - → RQ1: Has political violence increased recently?
  - → RQ2: Political violence now more likely by individual young men?
  - → RQ3: Is online radicalization the cause of this?
- Theoretical argument, hypotheses, expectations, etc

Good RQs, in brief

1. Empirically **answerable**
   - $\rightarrow$ i.e. you can answer it with data
2. Theoretically **relevant**
   - $\rightarrow$ i.e. it helps you learn something about your theory/argument

Good RQs, more in detail

1. Consider potential results of the analyses
   - → if you found X, does that answer the question? causally?
   - → example: are kids who play videogames often more aggresive?
   - → does that inform a theory on the aggresiveness effect of VGs? (you can even try to do a better RQ without causal ID)
2. Is it feasible?
   - → do you have the data? is it possible to do it? (e.g. re-offenders)
   - → also: is there any design or strategy to answer it?
3. Keeping it simple and narrow
   - → what are the causes of economic underdevelopment? vs. does exposure to natural disasters hinders economic development?

Good RQs, more in detail

This:

→ Has political violence increased recently?

→ Political violence now more likely by individual young men?

→ Is online radicalization the cause of this?

vs this:

→ What's the link between online media and political violence? (too broad)

→ Is online media the cause of the increase of political violence among young men? (many questions in one)

→ Why does online media drive political violence? (??)

Example on generating RQs

- Couple things to remember:
  - → RQs are often the link between theory and empirics
  - → So they already suggest which variation to look at

- Imagine you have the following argument:

  How good students do at school depends more on the peers they are surrounded by than on the quality of the teaching they receive

- Which RQ could let us test this?

Stories, RQs, and theories

- There're no exclusive definitions of 'stories' and theories
- It's just about getting to a sufficient level of abstraction
- Often, you start with a story or example, and then you move up the ladder for both theory and RQs until you get to a general theory tested with a RQ

Generating theories

- No recipe for this, everyone generates theories *all the time*
- Usually it refers to an analytical argument that explains something
  → It could also be a descriptive or predictive theory, but even in those cases there's probably an explanation underneath
- Developed inductively, from descriptive data to general explanations
- My advice: if you can't tell a story out of the theory, you're not there yet (i.e. need to be able to travel from/to abstraction)

- **Q:** How to identify a **good theory**?

Evaluating theories

1. Simple
2. Internally coherent and able to explain variation
3. Testable

Example (of the whole process)

Example

- That's some descriptive evidence that could inspire an anecdote
- The **anecdotal argument** (think of a story)
  - → My friend John who went on Erasmus has more money than my other friend who couldn't go, and also, John managed to get a job because his father is partner at a local firm
- The **research question**
  - → Is there a causal effect of Erasmus on labor market early success? Is the effect mediated by household income?
- The 'proper' **theory**
  - → Going on Erasmus does not have any causal effect on getting a first job, the relationship is explained by the confounding effect of income
  - → Or: Positive effect among high-income students because they have access to informal networks where this experience is valued
- The **hypotheses**?

Hypotheses

- Hypothesis is just a very formal term for empirical expectations
  - → Which essentially means being able to say what you expect to see given a theory

    And ideally, knowing what you need to see in order to discredit the theory

- Imagine that I have the theory that my knee hurts when I do sports on cold days
  - → Simplifying it, we have a 2x2 situation:

Hypotheses: what would you expect to observe?

	Cold day	*Hot day*
Run	Pain	Not pain
Didn't run	Not pain	Not pain

What if you observe this? **New theory? Test?**

	Cold day	*Hot day*
Run	Pain	Not pain
Didn't run	Pain	Pain

Computational methods and theory

- Limitations of *data mining*
- Focus on the what rather that on the why
- Problems with machine learning
  - $\rightarrow$ Example of predicting ice cream sales

Why bother about theory?

CIVIL UNREST

Predicting Civil Conflict: What Machine Learning Can Tell Us

Computer programs can be used as early warning systems, allowing the global community to act before violence erupts.

Mechanisms briefly

- A **mechanism** is basically the **how** (or why) of a relationship
- e.g., we know that the flu gives us fever
  - $\rightarrow$ flu > fever
- What's the mechanism?
  - $\rightarrow$ flu > *immune system detects infection* > Δ *body temp* > fever
- Good thing about mechanisms is that we can try to test them

Testing mechanisms (≈ sub-research questions)

- Let's go back to the Erasmus example
- If our theory is that the effect of going on Eramus is higher for high-income students due to their access to social networks, **what's the mechanism?**
- And how could we try to **test it**?
  - → (Think about the sub-RQs)

Recap

- Questions on theory, RQ, or mechanisms?

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Example: Wartime civilian deaths

Paper discussion

Concepts

- What are **concepts**?
- basically they are the **building blocks of analytical arguments**, so part of the theoretical framework

- Some times they are easy (income), but in many cases we have to think about them
  - $\rightarrow$ *Household* income? What's considered a household?
  - $\rightarrow$ More problematic: Political violence? Ideology? Democracy?
- Also minor point: *concept* $\neq$ *term*
  - $\rightarrow$ Think about the labels we use to refer to some particular concept, e.g. authoritarian regime (i.e. dictatorship), rationality, etc

Concepts types

- No fixed categories, but some people talk of:

1. **Rule-based** (definition)
   → e.g. what are the rules we could use to define a household?
2. **Ideal types** (or family resemblance?)
   → How do household look like? Can we intuitively identify them?

- Rather that two exclusive types of concepts, they are two ways to think about them which are usually useful in improving concepts

- Try: concept of **political violence**

Operationalization

- To translate abstract concepts into concrete stuff we can observe and potentially measure
- Operationalize $\neq$ measure
  - $\rightarrow$ The fact that you can think of a concept in concrete terms does not mean you can always measure it easily
  - $\rightarrow$ Remember the algorithm that models rate of re-offenders
  - $\rightarrow$ Ideology of Twitter users? easy to op, hard to measure **(*)**
- More like thinking of real-world attributes that map the conceptual dimensions we think about
  - $\rightarrow$ *concept*: war intensity; *operationalization*: number of battle deaths

Importance

- Might seem like something too abstract to care about (especially for computational social science), but it is actually not
- A **huge** part of good quantiative work relies on improving current concepts and their operationalization (which often leads to new ways of measuring them)
  - $\rightarrow$ today's paper ('Roads to rule') is a good example of this

Example

- Say we have a question about some x cause of civil war outbreak
- That's two concepts we are actually talking about:

1. **Civil war**
2. **Outbreak**

- How could we define them? And operationalize them?

More examples

- Let's say we think: "Political violence is not related to social media but to populist rhetoric"
- What is **populism**? How can we operationalize it?
  - $\rightarrow$ e.g. how could we code a list of *populist* political parties? or leaders?

(re: moving concepts up and down different levels of analysis)

Roadmap

Measurement issues

Three things to keep in mind:

1. Measuring what you really want to measure
   - $\rightarrow$ Careful with the use of proxies
2. Keep in mind what units you're not observing
   - $\rightarrow$ Missing data, sampling bias
3. Choosing the right unit of analysis
   - $\rightarrow$ Depends on the theory

Measurement issue #1

Measuring the right stuff

Measuring the right stuff

- Not a lot to say here, other than to **pay attention**
- We normally look at one variable superficially without thinking about how it was created
- Take a closer look: how was it exactly measured?
  - → Survey wordings
  - → Coding issues (e.g. level of democracy)
  - → Type of raw data used
- More importantly, are there **biases related to our question**?

(A few strategies to measure stuff not directly observable)

Example: we want to measure the ideological or policy positions of political parties

- **Expert surveys**
  - → You send questionnaires to experts who then reply, aggregate using average or similar
- **Coding written texts**
  - → Manifesto project, but also others based on NLP
- **Observing roll call voting**
  - → Voteview project

- All these point to slightly different concepts or operationalizations
- We'll see a different strategy based on *latent variables* in a moment

Measuring the right stuff: Example

- You are doing research on whether discrimination of minorities has a negative effect on overall economic performance of a country
- You find a dataset that lists all minorities in a given country and gives them a yearly score of discrimination from 0 to 10
  - → In the codebook says that discrimination is conceptualized as 'unequal access to state power, which ranges from actual, active discrimination (including mass violence perpetrated against members of the minority group) to lack of access to key political positions in the central government'
- You also learn that the dataset was coded through **expert surveys**, sending a questionnaire to 2–3 researchers from each country
- **What do you think?**

Measuring the right stuff: Example

- Now imagine you use the same dataset to analyze whether more extreme forms of discrimination make violence against minorities more likely
  - → You take the violence data from another dataset that e.g. codes actual violence events from newspapers
- You find a *positive relationship* in the results
- Thoughts?
- (Go back to previous definition)
- Another issue: difference depending on whether you do *within-country or between-country comparisons*

Measuring the right stuff: Another example

- Recent debate on **democratic backsliding**
- Problem: how do we measure democracy?
- Available data: many international datasets on democracy rely on subjetive **expert judgement**

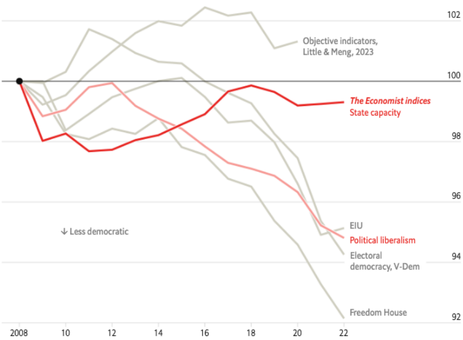
Graphic detail | Measure for measure

Global democratic backsliding seems real, even if it is hard to measure

Our analysis highlights two measures of governance that have diverged in recent years

Sep 12th 2023

Global democracy indices, 2008=100



Comment and Controversy

Measuring Democratic Backsliding

Andrew T. Little, *University of California, Berkeley, USA*
Anne Meng, *University of Virginia, USA*

ABSTRACT Despite the general narrative that the world is in a period of democratic decline, there have been surprisingly few empirical studies that assess whether this is systematically true. Most existing studies of global backsliding are based largely if not entirely on subjective indicators that rely on expert coder judgment. Our study surveys objective indicators of democracy (e.g., incumbent performance in elections) and finds little evidence of global democratic decline during the past decade. To explain the discrepancy in trends between expert-coded and objective indicators, we consider the role of coder bias and leaders strategically using more subtle undemocratic action. Although we cannot rule out the possibility that the world is becoming less democratic exclusively in ways that require subjective judgment to detect, this claim is not justified by existing evidence.

Democratic backsliding

- 'Objective' and 'subjective' operationalization & measurement
- 'Objective' measures usually rely on a *minimalist* conceptualization of democracy
  - → e.g. celebration of contested elections
- 'Subjective' measures tap into *maximalist* definitions of democracy that incorporate more dimensions
  - → regular, contested elections; but also rule of law, participation, media, etc
- Problem: Autocrats are usually pretty creative

Some support objective measures

- One example from the ACLP (Alvarez, Cheibub, Limongi, Przeworski) Democracy and Dictatorship Dataset
- Coding democracy based on four objective, observable rules:

1. The chief executive must be chosen by popular election or by a body that was itself popularly elected.
2. The legislature must be popularly elected.
3. There must be more than one party competing in the elections.
4. An alternation in power under electoral rules identical to the ones that brought the incumbent to office must have taken place.

Others prefer subjective ones (V-Dem)

1. Even if using expert surveys, you can take some measures
   - → Aim for *replicability*
   - → Incorporate measures of uncertainty
   - → Build it differently: e.g. incorporate different dimensions, use an ordinal scale, aggregate differently, etc
2. 'Objective measures' are not that objective
   - → Botswana example, and systematic downward bias against young democracies with economic growth
   - → e.g. how do you detect fraud? election forensics methods (based on distribution) can be incorporated by autocrats in later elections

- To know more: `v-dem.net/media/publications/wp_140.pdf`

Proxies

- A **proxy variable** is a variable that we use to substitute another variable we cannot observe or measure
- This is a matter of creativity, but the important thing is to think about **potential biases**. Think about these examples:
  - → Economic development and nightlight emissions
  - → ...or at the individual level, by having a TV
  - → Migration (in early 20thC Spain) proxied by excess male in censuses
  - → Effect of listening to the radio, proxied by transport mode (car)

Latent variables

- Some concepts are just not directly observable
  - $\rightarrow$ (or very expensive / unfeasible to do so)
- Another option is to create the variable out of *several other* observables
- This is sometimes called **latent variable approach**

Latent variables

- Let's look at one example: imagine you want to do research on whether left-wing or right-wing people tweet differently (or some other outcome, e.g. echo chambers idea)
- It's easy to get data on the outcome variable

  (Tweet content, frequency, ...)
- But **how do you code ideology?**
  - → Some people have done it focusing only on a subset, e.g. politicians, for which you have information (problem of selection)
  - → Of even some others have linked survey data to Twitter activity, asking for consent (problem of cost, non-response)

Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data

Pablo Barberá

Wilf Family Department of Politics, New York University, 19 W 4th Street, 2nd Floor, New York, NY 10012.
e-mail: pablo.barbera@nyu.edu

Edited by R. Michael Alvarez

Politicians and citizens increasingly engage in political conversations on social media outlets such as Twitter. In this article, I show that the structure of the social networks in which they are embedded can be a source of information about their ideological positions. Under the assumption that social networks are homophilic, I develop a Bayesian Spatial Following model that considers ideology as a latent variable, whose value can be inferred by examining which politics actors each user is following. This method allows us to estimate ideology for more actors than any existing alternative, at any point in time and across many polities. I apply this method to estimate ideal points for a large sample of both elite and mass public Twitter users in the United States and five European countries. The estimated positions of legislators and political parties replicate conventional measures of ideology. The method is also able to successfully classify individuals who state their political preferences publicly and a sample of users matched with their party registration records. To illustrate the potential contribution of these estimates, I examine the extent to which online behavior during the 2012 US presidential election campaign is clustered along ideological lines.

2 Ideal Point Estimation Using Twitter Data

2.1 *Assumptions*

In this article, I demonstrate that valid ideal point estimates of individual Twitter users and political actors with a Twitter account can be derived from the structure of the "following" links between these two sets of users. In order to do so, I develop a Bayesian spatial model of Twitter users' following behavior.

The key assumption of this model is that Twitter users prefer to follow politicians whose positions on the latent ideological dimension are similar to theirs. This assumption is equivalent to that of spatial voting models (see, e.g., Enelow and Hinich 1984). I consider following decisions to be costly signals about users' perceptions of both their ideological location and that of political accounts. Such cost can take two forms. If the content of the messages users are exposed to as a

Validating measure

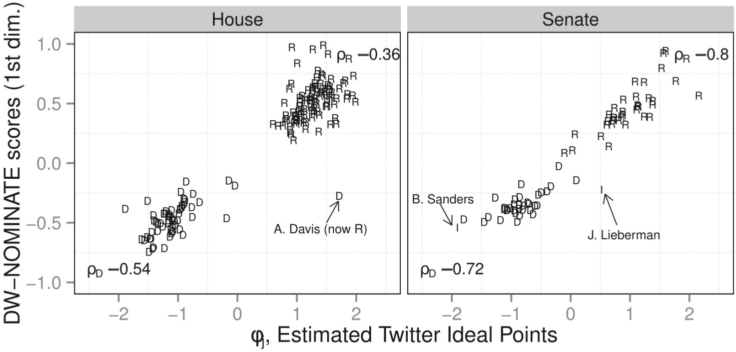


Fig. 1 Ideal point estimates for members of US Congress.

Measurement issue #2

What units are you not measuring?

Beyond our observations: missing data

- **Missing data** is when we don't have data for some observations
  → often more important that it looks, important to understand if it's biasing our analyses

1. Missing **completely at random**
   → No problem, random observations are missing
   (Probably not very often)
2. Missing **at random**
   → One variable explains whether obs are missing or not, but it's not related to our question
3. Missing **not at random**
   → The variable that explains 'missingness' is key to our question

Beyond our observations: missing data

- You want to know if some habit reduces post-surgery recovery time
- Data you have was collected as patients check in a hospital, but some observations are missing. Why?

1. A random group of forms were lost $\Rightarrow$ MCAR
   $\rightarrow$ (although...)
2. Some admin worker always forgets to ask filling the form $\Rightarrow$ MAR
   $\rightarrow$ (although...)
3. Medical emergencies do not fill up forms $\Rightarrow$ MNAR

Beyond our observations: sampling bias

- **Sampling bias** could be thought of as missing data or, rather, as a controlling variable we're indirectly including
- Easy case: we're dealing with a pre-designed sample that might have some biases
  - → Online survey and +65
- More easy to miss: there is an 'invisible variable' determining which observations we have or not
  - → e.g. social media consent-based experiment
- We'll talk more about how this affects inference
  - → Collider bias example?

Measurement issue #3

At which level should we measure?

Unit of analyses

- Level at which we have our observations
- Deeply related to the variables we have
  - $\rightarrow$ Even though not all variables have to/can be measured at the same level
  - $\rightarrow$ e.g. individual-level data and household income
- Most important thing: we need to **choose the right unit of analyses** depending on the theory (and mechanism) we are testing
  - $\rightarrow$ Getting the variation that matters (example of online bookings again)

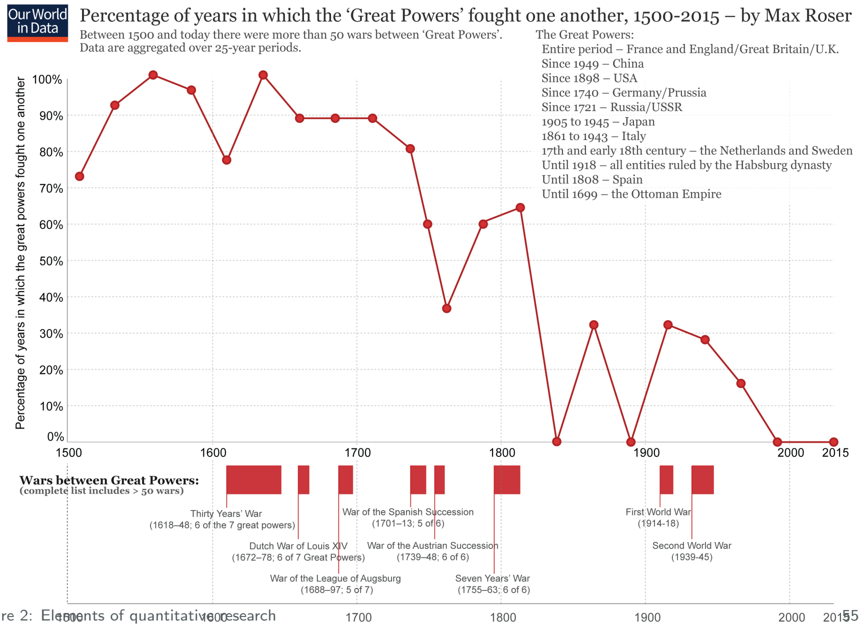
Theories, hypotheses, and measurement

- Let's say I want to explain the effect of school choice on future salaries
- My argument is: going to private schools leads to higher salaries in the future because increased resources lead to better educational attainment through lower teacher/pupil ratio, which signals individuals as more skillful in the labour market, explaining higher salaries

- Hypotheses?
- Testing the relationship and the mechanism?

  And alternative explanations?

Another example

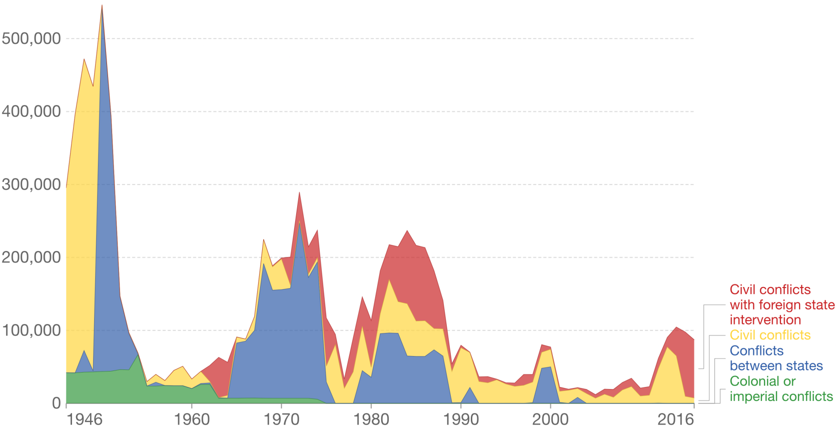


Another example

Battle-related deaths in state-based conflicts since 1946, 1946 to 2016

Only conflicts in which at least one party was the government of a state and which generated more than 25 battle-related deaths are included. The data refer to direct violent deaths. Deaths due to disease or famine caused by conflict are excluded. Extra-judicial killings in custody are also excluded.

Our World in Data



Source: UCDP/PRIO

Note: The war categories paraphrase UCDP/PRIO's technical definitions of 'Extrasystemic', 'Internal', 'Internationalised internal' and 'Interstate' respectively. In a small number of cases where wars were ascribed more than one type, deaths have been apportioned evenly to each type.

Another example

- Now, let's say my theory is that inter-state war has declined because democratic countries are less likely to go to war because they face higher domestic costs for waging wars

- Hypotheses? Testing the mechanism? Unit of analyses? Measurement? Alternative explanations?

Another example

- What if I say that it is because democracies do not fight *each other*, as they have shared interests in the international system and shared conflict resolution mechanisms?
- And if I say that democratic countries face higher costs when fighting another democracy, but not otherwise?
- What should I observe in each case? At different levels? How to measure?

Complexity of the social world and micro/macro

- Why do mass protests emerge?

Levels of explanation

Can you think of...?

- Macro-level mechanisms
- Micro-level mechanisms

- What's the point of macro-level explanations, actually?

Roadmap

Describing variables

- What is a **variable**?
- Types?
  - → Continuous
  - → Count
  - → Ordinal
  - → Categorical (binary)
  - → Qualitative (* are really a variable?)
- Why does it matter?
- Conceptual meaning vs statistical meaning

Describing variables

- Main idea: you are describing the variable distribution (i.e. how the frequency of values looks like)
- You probably know this from basic statistics
  → In practice, the measures of distribution do not matter so much
- But one important thing: we are talking about **real-world observations**, so before you do anything (analyses, etc), do look at them
  → At least, **plot the main variables**
  → Is it coherent with the **theoretical** or **expected distribution**?

Describing variables

- Also, sometimes the distribution is important to think about actual effect sizes, so it's good to summarize variables (mean, SD, IQR...)
  - → Maybe this makes sense if you've learn logistic regression?
  - → We'll talk more tomorrow about the concept of average effect in causality
- In a normal distribution, there's probably not much to say
- But what if a key independent variable has a bimodal distribution? What does this say about the **causal mechanism?**
  - → e.g. think about the effect of income on X in two societies: one is extremely unequal and the other is normally distributed

Describing relationships

- What is a **relationship**?
- Essentially that as you know about the values of one variable, you learn about the values of the other variables

  e.g. a *negative* relationship means that you know that higher values in x imply lower values in y

Imagine you have a small car, and a friend of yours is coming and is bringing along his two kids. Concerned about space, you ask '*how old are they?*' And the answer is: '*They're 6 and 2*.'

$\rightarrow$ What do you imagine about size?

$\rightarrow$ Now imagine you ask '*are they blonde, red-haired, or brown-haired?*'

Is description useful?

- There's no need to always ask about the *why*
- A description-based project could also be very complex: e.g. operationalizing and measuring
- For example:
  - → Some theories are just descriptive: six degrees of separation
  - → Some questions: has political violence/autocracy increased?
  - → Prediction can also be a form of description

Univariate description

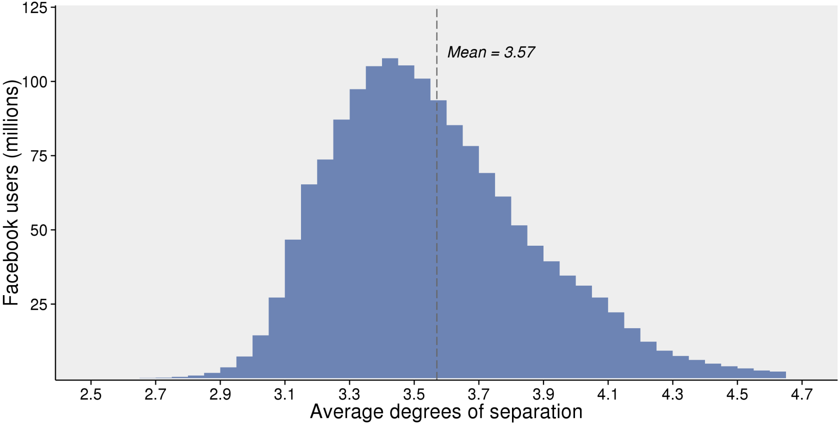


Fig 1. Estimated average degrees of separation between all people on FB.

(https://research.facebook.com/blog/2016/2/three-and-a-half-degrees-of-separation/)

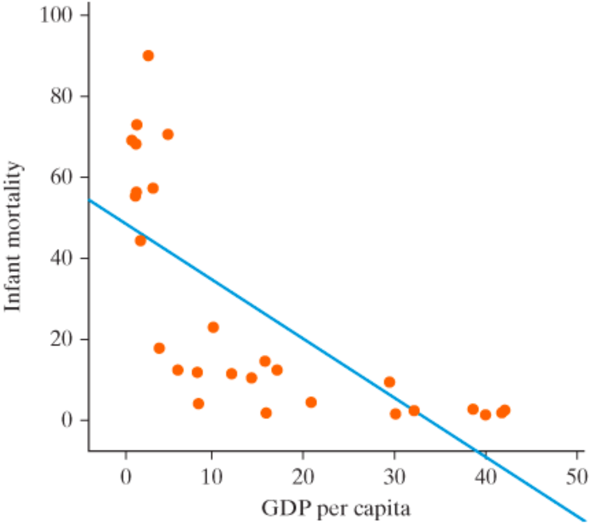
The original 'theory'

I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice. fill in the names.

Six Degrees of Separation, John Guare

- Do we have an answer?

Bivariate relationship



Bivariate relationships

- What do we use this for?
- Essentially, we are trying to detect whether two variables are dependent
  - $\rightarrow$ In other words, it's about conditional values: $E(Y|X)$
- Example graph about infant mortality:

  $E(IM|GDPpc = 1000)$?

  $E(IM|GDPpc = 30000)$?

Bivariate relationships

- This is what statistics is about, and only this
- Even if it can get complicated: non-linear relationships, multivariate dependencies, etc

Bivariate relationships

nature > nature communications > articles > article

Article | Open Access | Published: 20 August 2018

Sequences of purchases in credit card data reveal lifestyles in urban populations

Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla & Marta C. González

Nature Communications **9**, Article number: 3330 (2018) | Cite this article

This article has been updated

Abstract

Zipf-like distributions characterize a wide set of phenomena in physics, biology, economics, and social sciences. In human activities, Zipf's law describes, for example, the frequency of appearance of words in a text or the purchase types in shopping patterns. In the latter, the uneven distribution of transaction types is bound with the temporal sequences of purchases of individual choices. In this work, we define a framework using a text compression technique

Bivariate relationships

RESEARCH ARTICLE

Faces in the crowd: Twitter as alternative to protest surveys

Christopher Barrie[1☯]*, **Arun Frey**[2,3☯]

1 School of Social & Political Science, University of Edinburgh, Scotland, United Kingdom, **2** Department of Sociology, University of Oxford, England, United Kingdom, **3** Leverhulme Centre for Demographic Science, Oxford, England, United Kingdom

☯ All these authors contributed equally to this work and are listed in alphabetical order.
* christopher.barrie@ed.ac.uk

Abstract

Who goes to protests? To answer this question, existing research has relied either on retrospective surveys of populations or in-protest surveys of participants. Both techniques are prohibitively costly and face logistical and methodological constraints. In this article, we investigate the possibility of surveying protests using Twitter. We propose two techniques for sampling protestors on the ground from digital traces and estimate the demographic and ideological composition of ten protestor crowds using multidimensional scaling and machine-learning techniques. We test the accuracy of our estimates by comparing to two in-protest surveys from the 2017 Women's March in Washington, D.C. Results show that our Twitter sampling techniques are superior to hashtag sampling alone. They also approximate the ideology and gender distributions derived from on-the-ground surveys, albeit with some bias, but fail to retrieve accurate age group estimates. We conclude that online samples are

Describing relationships

- When we find a conditional relationship, we often say that X *explains* Y
- But these statistical relationships do *not* tell us anything about cause and effect, only about conditional means (or $E(Y|X)$, or conditional conditional means if we also control for Z)
- We need another strategy to understand *why*

Roadmap

Theories and research questions

Concepts and operationalization

Measurement

Description

Example: Wartime civilian deaths

Paper discussion

Practical example

CNN World | Africa | Americas | Asia | Australia | China | Europe | India | Middle East | More

Ukraine says it has identified a Russian commander accused of Bucha atrocities

By Mariya Knight and Radina Gigova, CNN

2 minute read · Published 1:04 AM EDT, Sat September 2, 2023

Practical example

- You want to test an argument about **wartime civilian deaths:**
  → The intuition you have is that civilians will be more likely to be treated well (and not killed) by armed groups during civil wars when they need their resources (e.g. labor) to survive

- Clean up the theory, decide on the main concepts
- Develop different RQ at different levels
- How can we measure the main concepts? Variables?
- What answers could we get from the data?
  → Are we learning something about our theory?

Roadmap

Journal of Conflict Resolution
2021, Vol. 65(2-3) 563-590

DOI: 10.1177/0022002720963674
journals.sagepub.com/home/jcr

Roads to Rule, Roads to Rebel: Relational State Capacity and Conflict in Africa

Carl Müller-Crepon[1], Philipp Hunziker[2], and Lars-Erik Cederman[3]

Abstract

Weak state capacity is one of the most important explanations of civil conflict. Yet, current conceptualizations of state capacity typically focus only on the state while ignoring the relational nature of armed conflict. We argue that opportunities for conflict arise where relational state capacity is low, that is, where the state has less control over its subjects than its potential challengers. This occurs in ethnic groups that are poorly accessible from the state capital, but are internally highly interconnected. To test this argument, we digitize detailed African road maps and convert them into a road atlas akin to Google Maps. We measure the accessibility and internal connectedness of groups via travel times obtained from this atlas and simulate road networks for an instrumental variable design. Our findings suggest that low relational state capacity increases the risk of armed conflict in Africa.