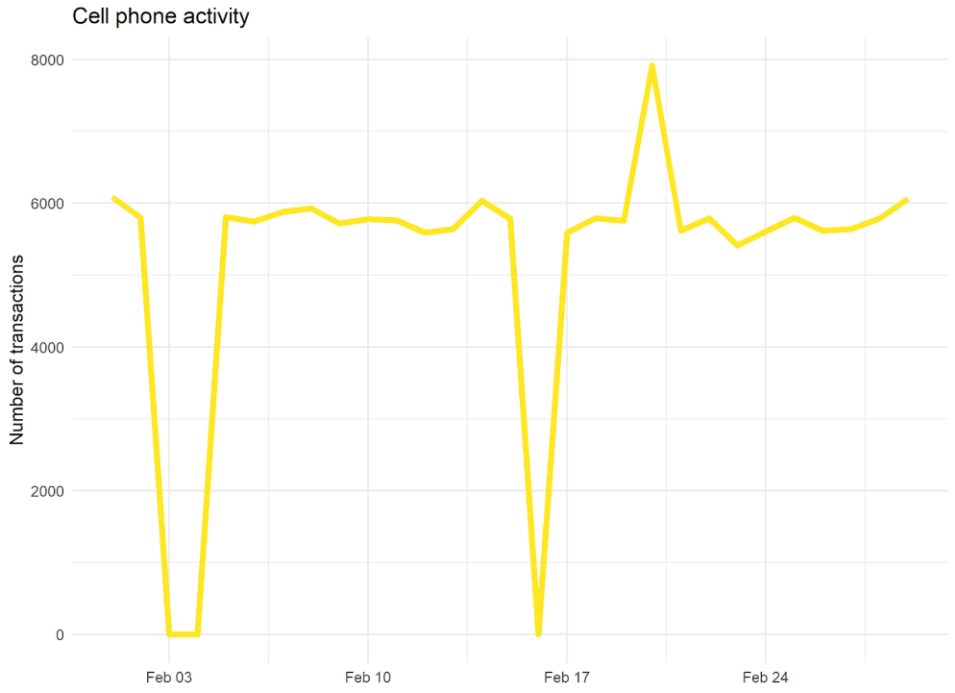
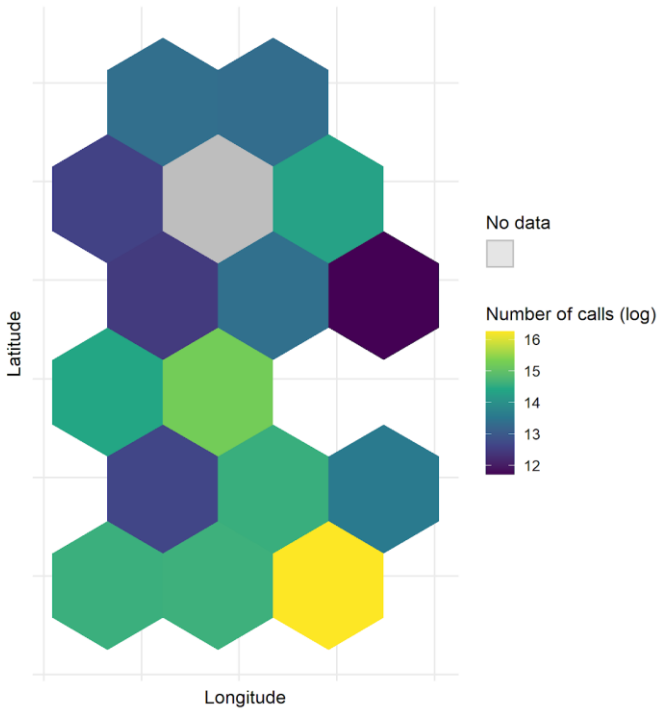
Big data quality matters

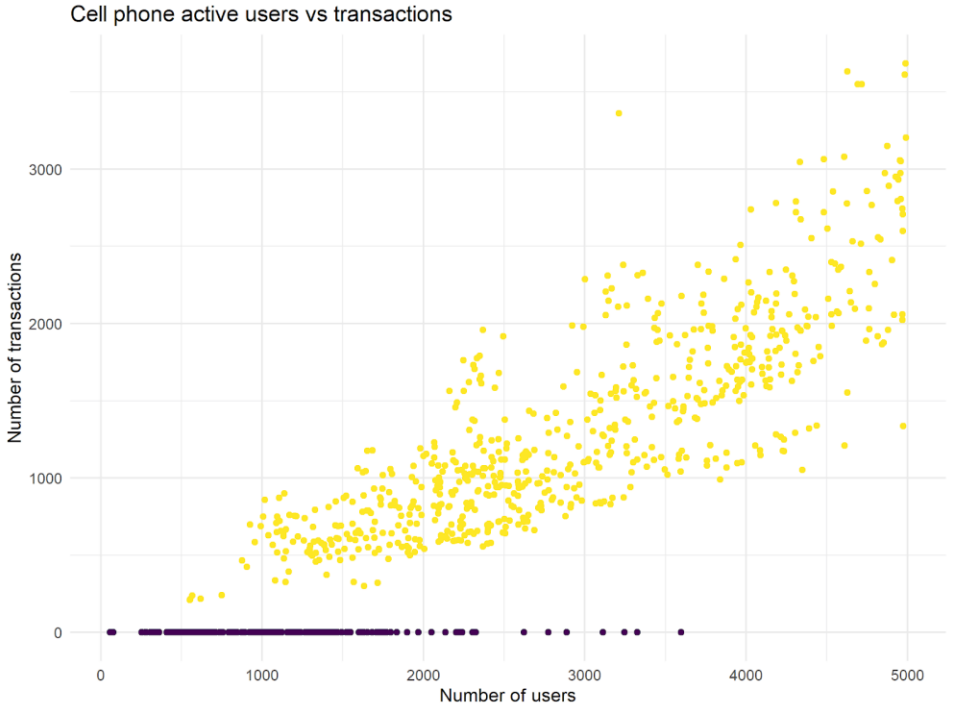
- Don't assume that big and real-time datasets are necessarily high quality
- Systematic data checks are a well-established best practice for surveys
- Data quality checks are as – if not more - important when using big data for development research, as data not purpose-built for research (unlike with surveys)
- New blogpost proposes 4 principles for data quality checks: https://blogs.worldbank.org/opendata/how-can-we-improve-quality-big-data-development-economics-research-experiences-traditional

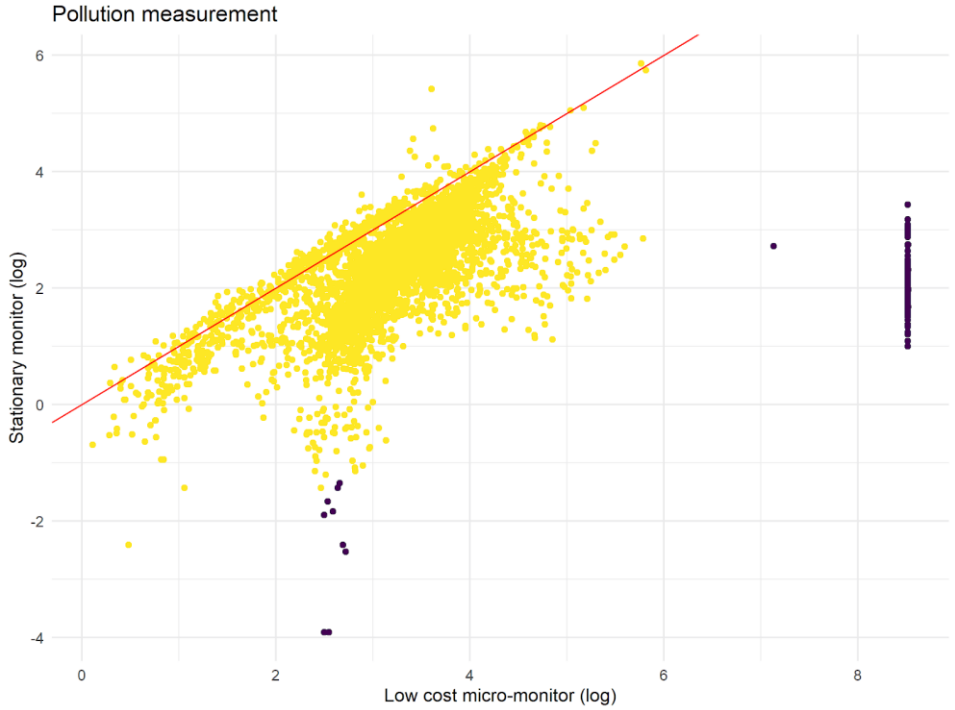
Data checks: completeness



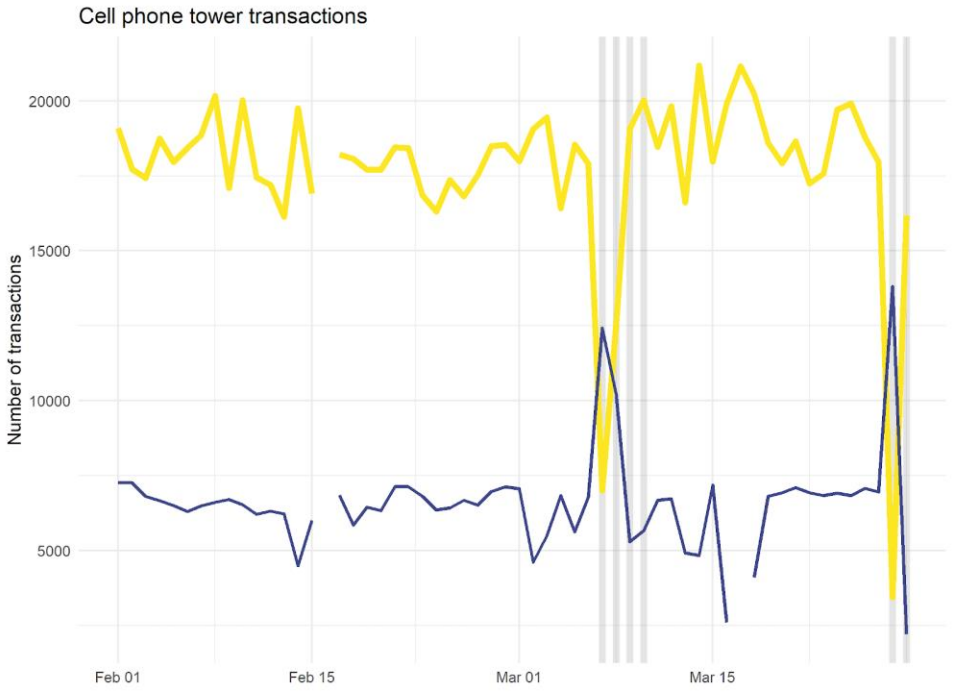


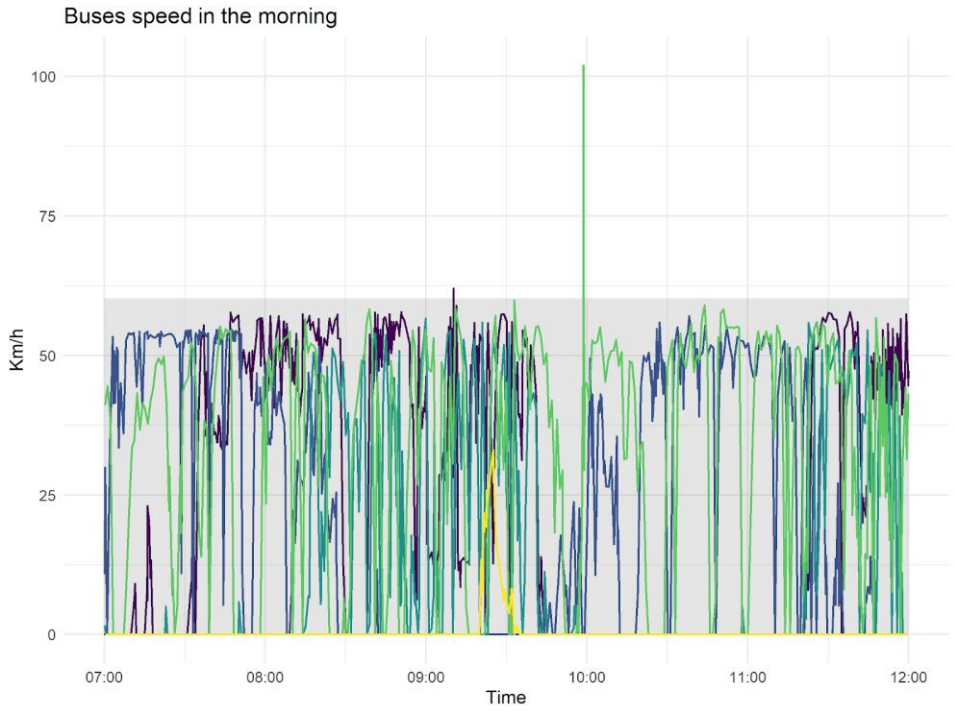
Data checks: consistency





Data checks: anomalous values





Data checks: real time

- Apply checks as soon as the data is received. The sooner issues are identified, the more likely they are to be fixed.

- *Example*: After setting up a microsensor to collect air quality data, our team checked the data coming in every day. One day we saw that no data was collected. Upon checking the sensor in person, it was discovered that after a power outage, the sensor did not properly begin collecting data again and it was necessary to reset. Only one day of data was lost as opposed to weeks if data checks had been done less frequently.

Reproducibility issues

- Analysis is resource-intensive, results can take a long time to reproduce.
- Slow connections and access to cloud compute resources are real constraints, especially in developing countries; computing costs difficult to anticipate/budget.
- Data often proprietary, owners may limit re-use or publication of the full data which inhibits reproducibility. Care needed in establishing licenses/terms-of-use.
- Privacy concerns! No well-established norms for de-identification. Risk/benefit calculus changes as more datasets released (and mosaiced).

DIME Analytics

Public Research Resources

- Development Research in Practice: the DIME Analytics Data Handbook
- DIME Wiki: a one-stop-shop for IE resources
- DIME Research Standards: research ethics, transparency, reproducibility, and open data
- Stata Visual Library, R Econ Visual Library : data visualizations with underlying source code
- ietoolkit and iefieldkit: Stata packages for impact evaluation analysis and data collection

Courses and training materials

- Manage Successful Impact Evaluations: flagship course for implementing IEs
- Manage Successful Impact Evaluation Surveys: virtual course for high quality data collection
- Research Assistant Onboarding : best practices for transparent, reproducible data work
- R for Advanced Stata users: hands-on training focused on basic data cleaning and analysis
- Measuring Development : annual conference