DataFirst
University of Cape Town • Private Bag X3
Rondebosch, 7701• South Africa
Tel: +27 21 650 5708 • Fax: +27 21 650 5709
Email: info@data1st.org
Website: http://www.datafirst.uct.ac.za

Report on Household Energy Data: Workshop 28 January 2016, University of Cape Town

This workshop was part of an ESRC-NRF International Centre Partnerships grant between UK Data Archive and DataFirst which is looking at a collaborative research infrastructure and data compilation project in the area of household energy data. A big focus of the project is anticipating and finding solutions for the challenges of 'big' and complex data in three main areas. The first is confidentiality, legal and ethical issues, particularly important in relation to administrative data and local level census and Demographic Surveillance Site data. The second relates to size and complexity, for example municipal datasets are of a size that social scientists in South Africa are not used to analysing, and the added challenges that comes with combining data from different data sources raises analysis issues of a different type. Third is surrounding data quality - missing data, data errors and uncertain or unknown provenance need to be identified and dealt with.

Professor Martin Wittenberg, Economist and Acting Director of DataFirst at University of Cape Town (UCT) introduced the focus of the workshop explaining that it was very much to kick start some knowledge engagement on these matters for social scientists in South Africa; and to brainstorm that might help lead to action and tangible research projects. A similar workshop had already been held with researchers in the UK in December 2015. One key aim was to identify potential data sources and research projects that could feature as demonstrators for the project. Participants came from various sectors spanning a good mix of energy researchers, economists and health researchers and national statisticians. These included: the UCT Energy Research Centre, the African Climate and Development Initiative, Statistics South Africa, the School of Economics, the Jameel Poverty Action Lab Africa (JPAL), the MRC/Wits University Rural Public Health and Health Transitions Research Unit (Agincourt) and Dartmouth University. Martin alluded to some recent high profile visitors to Cape Town, including the UK cricket team, Mumford and Sons, and now the UK Data Archive, surely the 'rock stars of data'!

The morning session comprised three talks on active energy research projects.

Municipal data

Grant Smith spoke about his and his colleague, Martine Visser's, work on the use of municipal electricity data derived from consumption data collected by municipalities based on billing information. Using prepaid vending data from some 600,000 customers (75 % of Cape Town customers from 2014 and some data dating back to 1993) with some detail on tariffs over time, billing records from 2005 for credit electricity, they are undertaking research on residential prepaid customers from 2004, and had already begun to analyse patterns of consumption among poor and rich households. The data came directly from commercial SAP business warehouse records, largely as text documents, plus records of transactions and debt. Across different departments there are various versions of data and in different structures. At a large (for social science) scale of 370 GB and some 300.000 households with daily data, a lot of data extraction and cleaning work needed to be done to extract the data for processing.

Added to this data were property values (in quantiles), property subdivision history, especially over the last 10 years, and GIS locations (as the primary key for linkage). Evidence arising so far was that the poorest households are having far smaller transactions per month. Further research papers would be coming on the provision of free water and electricity, and household welfare.

Issues arising from this area of work centred around four main areas. First, data extraction challenges at scale; the statistical processing is quite RAM constraining and the team felt they needed to learn new languages like Python to help manipulate data. Second is messiness and structure of the data sources; some standardised data collection and management by cities would be beneficial and could also be useful for direct interventions, such as the influence of nudges on behaviour across different groups. Fortunately, the ZA POPI (like UK DPA) does have caveats built in for research like this.

Third is the confidentiality of the data. The current data sharing agreement with the City is limited at present to a single research project, but a broader licence would be hugely beneficial, such as ones operated under the controlled environment of DataFirst's Research Data Centre. Finally, they are concerned with how to link this type of administrative data to other data sources (such as to survey data, via geocoded data/GPS coordinates) to gain a finer grained picture of local electricity access and use. This matching capability is needed to be able to evaluate national programmes and to look at trying to measure the developmental impacts of improving electricity access. This early work on Cape Town municipal data would likely be extensible to other municipalities in South Africa.

Rural electrification

The second session looked at projects on rural electrification. The electrification of South Africa's rural areas has significantly increased access since the end of Apartheid, and research is addressing what, if anything has been the impact of this? Mark Collinson and Taryn Dinkelman (via Skype) spoke about the Wits/MRC Rural Public Health and Health Transitions Research Unit (Agincourt) project which has been monitoring demographic and health changes in the Agincourt area since 1992. Since 2000 the team have also monitored changes in infrastructure and household assets as well as labour market outcomes. They had amassed a valuable set of panel data dating back to the 1990's, which would be turned into a shareable resource.

Questions for future work include whether we can see the impact of the electricity roll-outs on household outcomes in this area and what additional data, such as on timing of the roll-outs, consumption information, could we bring to bear.

Following a networking lunch, in the afternoon the UK team outlined the work being done in the UK to help 'scale up' data activities for social scientists, in particular building a prototype service infrastructure for managing and making big data accessible. Louise Corti, Director of Collections Development at the UK Data Archive at the University of Essex spoke about the UK's role in the UK-ZA Centre partnership grant. The focus on providing use cases for supporting managing and researching questions on household energy data would help the UK develop and deliver an infrastructure to include technical, governance and security models for ingest and access to data. A blueprint of a working system could then be delivered to DataFirst. She was very pleased to be working with the Centre for Environmental Epidemiology (CEE) at University College London (UCL) to provide the domain research expertise. So far the UK Data Service has attracted a few active researchers who were keen to pursue projects making use of the big data environment being developed; in essence to see how it compared with more traditional methods, and what analytic and visualisation tools might be available. The work also paid attention to methods of ingesting and managing data from real time data sources, such as regular data capture from household energy

smart meters, and linking of data sources that is likely to point to disclosure risk: how is this managed in a big data environment? A second aim was to undertake in both countries an audit of energy data sources. Finally, the project would be creating instructional materials to be used for a weeklong summer school for upskilling social scientists for managing and analysing new and novel forms of data, to be held in UCT in February 2017.

Simon Elam of the CEE spoke about the role of the CEE in brokering access to and use of energy smart meter data or the UK, based on the UK government's plan to install a meter in every household in the UK by 2019. The work of his centre aimed to: improve the collection of empirical energy demand and related data; help combine and curate relevant data sets and help researchers, industry and policy makers use energy data and the results of data analysis via securing open access to data.

Simon also spoke of the 'Vulnerable Customers and Energy Efficiency (VCEE)'project currently in operation which sought to engage with 'fuel poor' so they could benefit from energy efficiency and demand side response. The project included an energy supply company supplier for all VCEE participants who would install electricity and gas smart meters and temperature loggers in participant homes.

Dr. Nathan Cunningham spoke about the cutting edge work being done by the UK Data Service current Big Data Network Support award to develop a demonstrator for a modern, fit-for-purpose research data infrastructure utilising current industry-standard Open Data Platform technology in an affordable and accessible way. This modern platform that takes its standards from the open communities and aims to enable ready discovery, linkage and visualisation of datasets of interest to the social scientist. A key issue here is the moving of processing to the data and not the other way around. The UK Data Service is designing a blueprint for a fully functional infrastructure that could, if realised, be transferred and taken up by UCT.

The final session of the day saw Martin Wittenberg and his PhD student, Tom Harris talk about the opportunities that combining local and national data could bring. While the projects discussed in the morning involved micro-studies in a local context looking at two very different contexts, much of the national policy discussion actually occurred on the basis of information supplied by national level datasets (the census, information from the General Household Surveys, Eskom data). The session questioned whether national level datasets were informative enough about local conditions, or could be used to improve our understanding of local conditions, and whether local information could be used to improve the quality of national data.

Tom spoke of his and Martin's detailed work on the post-Apartheid labour market series (PALMS) data set. This included a large-scale rescue of older smaller surveys from the national statistical institute, administrative and market research data sources that involved harmonisation and cleaning of some 28 surveys. Tricky issues were constructing reliable weights over time, and some work had been done by Martin on smoothing weights and flagging outliers. Critical here was the release by DataFirst of the code for cleaning the data, demonstrating excellent research transparency.

The wrap up session invited participants to think about connecting up data, research questions and forward-looking infrastructure. The cocktail reception invited further networking and discussion about the pragmatics of working together to test out the open data platform and explore new ways of working that utilised opportunities offered by the big data platform.