GCRO OCCASIONAL PAPER
NO. 20

Adventures in city data: An ethnographic story

DECEMBER 2022
Author: Shirley Robinson

GCRO | Gauteng City-Region Observatory

The GCRO comprises a partnership of:

ADVENTURES IN CITY DATA: AN ETHNOGRAPHIC STORY

Publishers: Gauteng City-Region Observatory (GCRO), a partnership of the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government and organised local government in Gauteng (SALGA).

DOI: https://doi.org/10.36634/RPLR3362
ISBN (XML): 978-1-990972-27-0
ISBN (Web pdf): 978-1-990972-28-7

Author: Shirley Robinson
Design: Breinstorm Brand Architects
Typesetting: Lumina Datamatics
Editor and production manager: Simon Chislett
Cover image: SDecoret / Adobe Stock

GCRO OCCASIONAL PAPER
NO. 20

Adventures in city data: An ethnographic story

DECEMBER 2022
Author: Shirley Robinson

Contents

Figures

Acronyms and abbreviations

API	application programming interface
CIT	company income tax
ERLN	Economies of Regions Learning Network
GTAC	Government Technical Advisory Centre
ISIC	International Standard Industrial Classification of All Economic Activities
NT-SDF	National Treasury Secure Data Facility
PAYE	pay-as-you-earn
RSC	Regional Service Council
SA-TIED	Southern Africa–Towards Inclusive Economic Development
SARS	South African Revenue Service
SIC	Standard Classification of All Economic Activity
StatsSA	Statistics South Africa
UIF	Unemployment Insurance Fund
UNU-WIDER	United Nations University–World Institute for Development Economics Research
VAT	value-added tax

About the author

Shirley Robinson is an economist who obtained her Master's in Business Science from the University of Cape Town. With 25 years of applied public policy experience (focused in particular on South African and Southern African/East African urban public policy, public finance management and the broader inclusive economic development agenda), she has also provided support on sustainable development financing to the European Commission on European Development Policy. Her expertise lies in leveraging the synergies between urban economic development, innovative financing, results-based management, and monitoring and evaluation in socio-economic development.

As a Long-Term Technical Advisory to the National Treasury's Government Technical Advisory Centre, Shirley led the work on the city economic data programme in support of the first phase of the National Treasury's Cities Support Programme. During the recent years that Shirley spent in the Netherlands as a Senior Consultant at Ecorys NL, *Adventures in city data*, initially written as a draft for an academic paper, then emerged as an ethnographic account sponsored by the GCRO. Shirley is now back in South Africa and working at the Western Cape Provincial Treasury, where she heads the Public Policy team in the Fiscal and Economic Services division.

Acknowledgements

Producing *Adventures in city data* has been a journey of note. I would like to acknowledge and give thanks to : Roland Hunter, who, as my project manager at the National Treasury's Cities Support Programme (CSP), gave me the space to dream, create, do and fly – for which I will be forever grateful; Kirsten Pearson, my amazingly creative, energetic and always supportive co-lead in the ERLN Technical Working Group on Data; David Savage for his insightful leadership of the CSP and the partnership with the World Bank on the Urbanisation Review of South Africa, which served as a catalyst for driving the city economic data agenda forward; all the partners and stakeholders of the ERLN and the data custodians who told me their stories of their data-world; Chris Hiddink and Atze Verkennis, my managers at Ecorys NL, who gave me space to write this paper while on my sojourn in the Netherlands; Rashid Seedat, Rob Moore, Graeme Götz and the GCRO team, who gave me strength to believe in myself and to tell the story; and Simon Chislett, my editor and publication manager, who made my words dance and the story emerge in colour and with dynamism.

What started out as a draft for an academic paper became a story, one that made me realise that it is the strength of relationships that unlocks the evidence underlying the data platforms. I have learned much about the data world through this journey. I have learned even more about myself and the journey I would like to follow going forward. Thanks to you all for accompanying me, and I am looking forward to going on future adventures with many of you.

Foreword

The Gauteng City-Region Observatory (GCRO) mainly publishes research by its own staff in its various publications series. However, on occasion it also commissions pieces from specialists in the field, or publishes work offered to it by professionals or scholars that add significantly to the body of knowledge needed to aid development in the Gauteng City-Region (GCR). Examples of the latter include McCarthy (2010) on acid mine drainage in the GCR, Mabin (2013) on the long histories of planning in the southern Transvaal, and Harrison (2020) on the histories of epidemics in Johannesburg.

Shirley Robinson approached the GCRO in January 2021 with a view to publishing work she had started as part of her PhD. We were intrigued. Initial discussions focused on whether we could publish preliminary outputs from her doctoral project using anonymised tax information to analyse and visualise economic activity in South Africa's major cities. We reasoned that the analysis could expand the available picture of the GCR's economic geography. But we were particularly excited when Shirley subsequently shared with us some 15 000 words of detailed recollections and reflections on the protracted effort inside government to assemble the systems needed to extract and geocode this tax data. Working with other committed officials and consultants over many years, Shirley had been at the heart of this effort. Her complex but compelling backstory behind the data – based on her lived personal experience, yet at the same time carefully depersonalised – spoke to the GCRO's own research in a number of ways.

First, it reminded us of the many constraints we had encountered in our own efforts to map the urban space economy of Gauteng. We had certainly seen gains over the years, especially with innovative remote sensing data able to discern changes in the location of industrial and commercial buildings (Naidoo, 2019; Götz and Todes, 2014). But there had also been many frustrations with inconclusive analyses, and dead ends in trying to procure good quality datasets. As we read Shirley's review of the challenges that occasioned her and her colleagues' efforts, we couldn't help but recall the many appeals from our government partners to furnish more granular analysis on aspects of the region's economy – for pictures of economic output per electoral ward to assist with spatial targeting of development projects; for a geographically and sectorally segmented sample frame for a firm survey on constraints to doing business; for an analysis of growth prospects in the township economy; and so on. These and many similar requests have been impossible to meet because of the despairing lack of reliable and detailed spatial economic data.

Second, Shirley's detailed account of what it took to piece together data systems inside the state resonated with the GCRO's own experiences at the coalface of producing primary data. Our experience is that those who work in sophisticated ways with large datasets, applying advanced statistical techniques or even artificial intelligence algorithms to reach conclusions, often have little if any interest in the processes by which the data they use are generated. They tend to take the validity of the data at face value, trusting mathematical values internal to their calculations to vouch for accuracy and statistical confidence. It is an entirely different kind of researcher who worries about whether survey fieldworkers have done their job well, records have been entered consistently in databases, or geocoding engines have been able to correctly recognise an address. The GCRO has learned the hard way through a decade of Quality of Life surveys that data assembly is a messy, human process, where rigour cannot be taken for granted, but must instead be hard won through painstaking quality control. When we read Shirley's narrative in early 2021 it struck a particular chord. We had just spent many months asking hard questions around the geolocation of COVID-19 cases from poor quality address data, in the process annoying everyone from modellers who sought to reach grand conclusions around COVID

hotspot patterns, to government officials relying on the maps projected on their 'war room' viewscreens to coordinate response teams. So Shirley's cogent account of how government set out to solve an address problem inherent in the currently available economic datasets, a problem which remained stubbornly unsolvable despite best efforts, echoed loudly for us.

Third, the narrative offered to us was from someone who had worked over a long period – from 2014 to 2019 – to solve a problem within government for the National Treasury's Cities Support Programme (CSP). It gave a unique perspective of the inside workings of the state apparatus. Most scholarly work on government is from the outside looking in, and much tends to be normatively laden with easy-to-hand paradigms and terms that are preconceived to find fault. In recent years, research in the GCRO's thematic focus area on 'government and governance' has turned increasingly to a different kind of analysis. This does not stop short at external policy critique but tries to dig below the surface to more rigorously excavate what is going on inside government institutions, and thereby understand the reasons for failure and the prospects for doing better. Shirley reveals for us, methodically and carefully, not as a breathless exposé, aspects of how government functions below the surface. This is an extended 'adventure' of finicky information-problems in the process of being figured out, of exacting attempts to overcome organisational disconnects, of genuine and sustained effort towards innovation in the interest of cooperative governance and, ultimately, evidence-based decision-making. This is not to say that it is written as a feel-good story. We do get a clear sense of how progress gets frustrated by the dull inevitability of competing or narrowly interpreted mandates, everyday disagreements over next steps, the exit of key stakeholders because of capacity limits, and delays in signing memorandums of understanding. As such, it affords a dispassionate but penetrating deep-ethnography of government at work that seemed to fit with our own aspirations for new research in our governance theme.

In May 2021, some two years after Shirley stopped working on the initiative, and four months after she first approached us, the National Treasury/CSP published a first set of *City Spatialised Economic Data Report*s (CSP, 2021) analysing and visualising the anonymised tax data. As noted in this Occasional Paper, these reports represent a huge milestone in light of all the work that had been put in over time. They did unfortunately take away any impetus for us to publish the Gauteng results from the data, since the three metropolitan municipalities were each well covered by their own report. But they refocused our minds on the value of the backstory behind the data, accentuating for us how important it was to surface what is not in view in the reports, but that resonated so strongly for us when we read Shirley's systematic recollections.

The GCRO wishes to thank Shirley Robinson for being keen to publish her 'adventure' with us. We also wish to acknowledge and thank the many who gave input into draft work over time, including: GCRO staff, either individually or at a brownbag seminar; a colleague of Shirley's who had worked closely with her during the initiative, who also affirmed the importance of the account; two blind peer reviewers; and our production manager Simon Chislett for his careful curating and copy-editing of the final text.

Abstract

South Africa is urbanising rapidly and its economic landscape is continuously changing as a consequence. In this context, city governments and urban scientists have long called for better access to city economic data. The National Treasury's Cities Support Programme (CSP) has reinforced this demand, insisting that disaggregated intra-city anonymised and geocoded economic data is critical for improving economic planning, performance and investment in South Africa's cities.

A wealth of economic data on tax payers is collected by the South African Revenue Service in the course of their operations. In addition to its bureaucratic purpose, economic data represents an enormous resource for a detailed understanding of the urban economy. Until recently, this resource has been underutilised because it was not available in an anonymised and geocoded form. At a practical level, however, the significant amount of energy and time required to access, clean and align administrative datasets to make them usable is not generally understood.

This GCRO Occasional Paper presents an ethnographic account of the decade-long journey in city economic data collation that contributed to the milestone publication of the *2021 City Spatial Economic Data Reports*. After observing the critical need for anonymised and geocoded economic administrative data in policy formulation and urban research, this paper examines the reasons for the limited availability of datasets that show the locations and economic activity of jobs and production at a disaggregated local level.

The paper details how the National Treasury's collaboration in 2016 with the World Bank to produce the Urbanisation Review of South Africa stimulated and directed the efforts of the Government Technical Advisory Committee and the Economies of Regions Learning Network to pursue official sources of city-level administrative data.

The paper goes on to recount subsequent National Treasury/CSP collaborations with Statistics South Africa, the South African Revenue Service and the Unemployment Insurance Fund to collect and collate anonymised and geocoded city economic data from sources other than national general surveys. Most importantly, these efforts focused on obtaining data at the disaggregated local level rather than at the aggregated 'head-office' level that is routinely obtained in the general surveys. Despite progress, these efforts were ultimately stymied due to practical and governance constraints.

Nevertheless, the paper narrates how, in a parallel process, these collaborations ultimately bore fruit in the establishment of a secure administrative data centre at the National Treasury that stores anonymised data which has been geocoded using postal codes. While this method is not ideal, it remains the best alternative given the lack of structured and standardised address data in the available tax datasets.

The paper concludes by reflecting on the insights from this ethnographic account that are critical for improving the integrity of the city spatial economic data resource and to enhance its use in credible, evidence-based urban analysis. First, these conclusions draw on broader institutional and public management concerns within the current governance environment on which steps to improve city spatial economic data will depend. Second, the paper points out that, despite the long journey travelled, the aggregated head-office conundrum and business classification uncertainty still remain. Solving these governance and data puzzles will unlock the incredible potential that such an evidence-based resource holds for creating a more just and equal society in South Africa, and beyond.

Photograph by James Oatway

Photograph by Carols Castilla / Shutterstock

Introduction

Introduction

Story premise

This GCRO Occasional Paper presents an action-research ethnographic account[1] of a decade-long journey towards making intra-city spatial and economic administrative data (hereafter, referred to as 'city economic data') more available and accessible for urban research and policy formulation.

South Africa's city governments and urban scientists have long been calling for better access to city economic data, especially data that is spatially referenced, to support economic development policy and programme formulation. The National Treasury's Cities Support Programme (CSP) reinforced this demand, insisting that disaggregated intra-city anonymised and geocoded economic data is critical to improving economic planning, performance and investment in South Africa's cities.

A wealth of economic data is collected by the South African Revenue Service (SARS) as a routine part of tax compliance by tax payers. This data also represents an enormous resource for a detailed understanding of the economy. Until recently, this data resource has not been utilised. However, in May 2021, the Natural Treasury released a series of *City Spatialised Economic Data Reports* (CSP, 2021) for each metropolitan municipality in the country. These reports, based on anonymised tax data, are a first foray into using such data for intelligence on the economy itself.

The CSP's collaboration with the World Bank in their Urbanisation Review of South Africa 2016–2018 provided the inciting incident that enabled the CSP to take the lead in finding a solution to the World Bank's requirement for disaggregated city economic data. The CSP joined the National Treasury's Economies of Regions Learning Network (ERLN) and their Technical Working Group on Data in a dedicated partnership led by the Government Technical Advisory Centre (GTAC).

The author of this narrative headed GTAC's support to the CSP between 2014 and 2019. Between 2016 and 2018, efforts focused on obtaining the specific data required for the World Bank's Urbanisation Review, which involved the collaboration of multiple stakeholders, including Statistics South Africa (StatsSA), SARS, the Unemployment Insurance Fund (UIF) and various departments, bodies and workstreams within the National Treasury itself. These efforts culminated in the National Treasury setting up a dedicated secure data resource facility by 2019. The *2021 City Spatialised Economic Data Reports* are based on the data generated through this facility. These reports represent a significant milestone, showing how anonymised and geocoded tax microdata can be used to undertake sharper, deeper and more granular spatial analysis at the intra-city level.

Plotline of the adventures

The next section of this paper explains the context and rationale behind the importance of city economic data and why there have historically been gaps in its collection. It affirms the broad consensus that these gaps need to be closed using innovative approaches to obtain more granular, disaggregated data at the local establishment level rather than aggregated data at the

1 The paper's ethnographic approach is based on the author's lengthy participation in the data collaboration process, and her interviews during this process with various public data custodians. This ethnographic account does not purport to spell out a research method; rather it narrates the processes that took place to bring about improvements in the collation of city spatial economic data using administrative data, and as such can be seen as 'action research'.

'head-office' level that is traditionally obtained from general surveys. This section highlights the increasing voices of cities themselves, advocating for better access to city economic data for urban planning and service delivery.

The third section describes the World Bank's undertaking of the Urbanisation Review of South Africa in collaboration with the National Treasury from 2016 to 2018, which further emphasised the need for disaggregated city economic data and provided the impetus for concerted efforts to obtain anonymised and geocoded data at the establishment level. This section describes the initial endeavour to devise an appropriate sampling frame that could enable the geocoding of data obtained from StatsSA's quarterly labour force and household surveys. It also details the obstacles that eventually stymied this approach.

The fourth section then presents the backstory to the *City Spatialised Economic Data Reports* in the series of collaborative efforts and learnings of various 'data pathfinders'. These visionaries worked tirelessly in different conversations and task groups over the years – principally within the National Treasury's CSP and in collaboration with SARS and the UIF – to unlock access to city economic data. The following section recounts how the data pathfinders' combined efforts to establish a secure city economic data resource eventually culminated in the unveiling of the National Treasury Secure Data Facility (NT-SDF) in 2019.

The final section draws the story to a close, raising key insights from this ethnographic account on how to design the next phase, as city economic data is increasingly needed for evidence-based urban economic analysis, planning and service delivery. These findings point, first, to the governance solutions required to staff the sector with qualified personnel and make public institutions more robust. Second, they highlight the need to continue to work on solving the conundrums impeding the collection of accurate anonymised and geocoded data, including the persistent head-office effect (where, as explained below, large firms sum and report their employment and output data as if it all occurred at their central head-offices, rather than their geographically dispersed branches), and the fact that industry classifications remain largely imputed rather than clearly categorised. Finally, the paper notes that while the journey it describes is uniquely South African, its lessons on the use of anonymised and geocoded administrative data have wider appeal, particularly for other developing economies in the Global South (ATAF, 2021).

Figure 1: The *2021 City Spatialised Economic Data Reports* for the Gauteng City-Region's three metros
SOURCE: CSP (2021)

Joburg street map by Rahman Hilmy Nugroho / Shutterstock

Importance of city economic data

Importance of city economic data

South Africa's urbanisation imperative

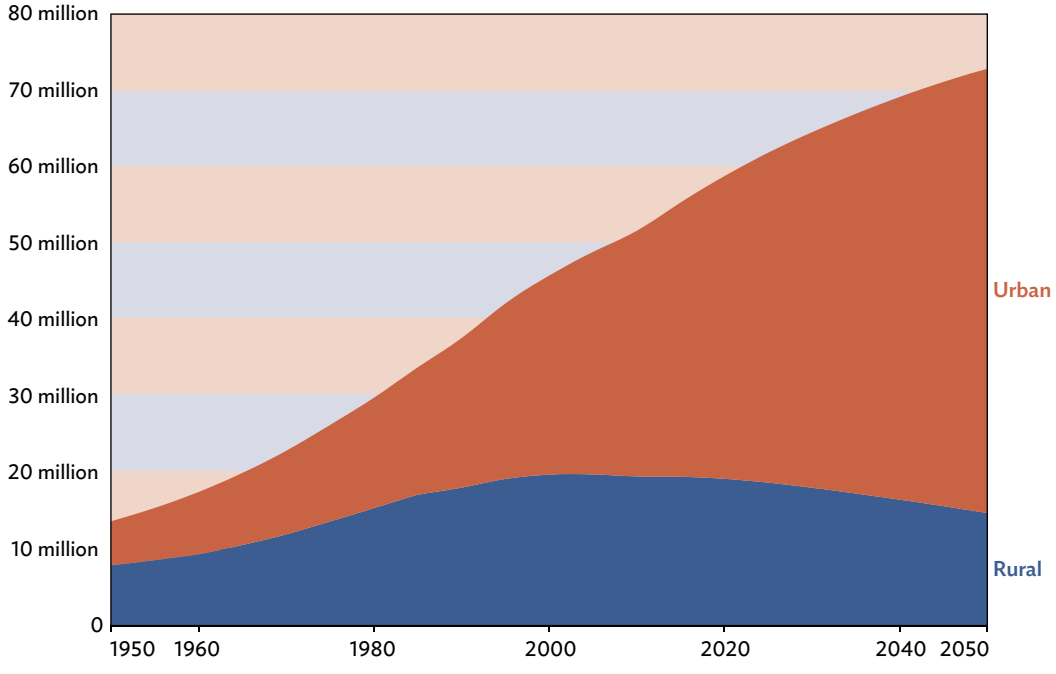
South Africa is urbanising rapidly and its economic landscape is continuously changing as a consequence. Massive migration since the demise of apartheid saw over a quarter of the country's population move in the five-year period after 1996 (Shilpi et al., 2018). Over two-thirds (66.4%) of South Africans now live in urban areas. The country's urban population will rise to 72.1% in the next decade and, by 2050, eight out of every ten South Africans will be urban citizens (UN DESA, 2018).

Evidence suggests that South Africans are highly mobile, moving mainly to the Gauteng conurbation (Johannesburg, Tshwane and Ekurhuleni) and the Cape Town area, where the greatest economic opportunities within South Africa are situated. These four metropolitan municipalities ('metros') received over half of South Africa's total population growth between 2001–2011 and generated just under half (47.3%) of employment growth over the same period (Turok and Borel-Saladin, 2014).

The scale and direction of urban migration is reshaping the contours of South Africa's urban economy in real time. Cities are not only growing outwardly but also densifying, since most urban

Figure 2: More than 50 million South Africans are projected to be living in urban areas by 2050

DATA SOURCE: Our World in Data; redrawn by Lumina Datamatics



growth takes place on already urbanised land that has to accommodate more people and their activities. Urban neighbourhoods, business districts, industrial parks and wastelands are transforming as they take on different usage, sounds, activities and atmospheres. Constant and consistent urban change is now South Africa's daily reality, particularly in a post-COVID-19 context. However, the spatial-economic impacts of urbanisation in South Africa are not fully articulated in evidence-based research given the limited spatial-economic analysis at the intra-city level.

Limited data's impact on urban research

Whereas empirical spatial-economic analysis is a common policy tool in many countries, it is relatively new in South Africa. To date, sub-national spatial-economic analysis has largely focused at the provincial or city-boundary level. There has been limited analysis at a more granular intra-city level, for instance on neighbourhood residential or investment profiles, because of the lack of current and publicly available data on economic activity and employment at an intra-city main-place level.[2]

Prior to the National Treasury's release of the city spatial-economic data resource in 2021, researchers had utilised either administrative data at the local level from the Regional Service Council (RSC) levy,[3] or proprietary data on economic activity and employment from private-sector data providers such as IHS Global Insight and Quantec (CSIR, 2020; Republic of South Africa, 2020; Turok and Borel-Saladin, 2013; Sinclair-Smith and Turok, 2012). However, a granular intra-city spatial dimension is crucial for meaningful analysis since people live, work and play in urban space, thereby shaping patterns of social and economic activity and contributing to the flows of goods and services that underpin a city's economic performance.

Urban economists debate whether it is either people moving to jobs or jobs relocating and people following them that is the catalyst of urban change. Storper (2013, p. 14) asserts that cities are the 'workshops of the world', and that major urban change will be driven by the location choices of firms and industries, with people as workers and consumers following suit. In confirmation, the 2018 World Economic Forum's *Future of Jobs Report* (WEF, 2018) draws attention to widespread changes in the global geography of value chains as technological advances accelerate transformation and drive firms' location decisions. This dynamic is already playing out in South Africa's rapid urbanisation as people relocate or are pulled towards greater economic opportunity in response to the spatial mismatches caused by apartheid's segregationist policies (Crankshaw, 2020; Shilpi et. al., 2018; Todes and Turok, 2018; Turok et al., 2017; Turok and Borel-Saladin, 2013).

Lamentably, South Africa's lack of publicly accessible city economic data sources means that urban policy-makers and researchers do not know exactly where firms and jobs are located. This is because the available research is either dated or utilises proprietary economic data that

2 'Main-place' refers to an official spatial layer with the names of cities, towns, townships and villages in South Africa. South Africa's statistical geographic hierarchy divides the country into nine provinces, and then further into eight metropolitan municipalities (largest cities) and 44 district municipalities. The district municipalities are divided into 226 local municipalities. At a lower level of geographic hierarchy, main-places are named locations determined by StatsSA that include towns, small cities, regions of large cities, or tribal areas. Sub-places are named locations determined by StatsSA that generally correspond to suburbs, villages or localities (StatsSA, 2011). As a case in point, the Centre for Affordable Housing Finance in Africa has recently published an analysis of eThekwini's residential property markets (CAFH, 2021) with a particular focus at the lower-end or township property market, using title deeds data purchased from Lightstone, a private company that provides data and analysis on the South African property market.

3 The RSC levy was a local business-turnover and payroll tax levied by metropolitan and district municipalities that was phased out in 2006 because it was considered economically and administratively inefficient.

still reflects the aggregated 'head-office effect' rather than providing data at the geographical establishment level.[4] However, spatial-economic data is critical for a myriad of reasons. Access to jobs has to date been a key driver of household decisions on where to live in a city's space economy. Knowing where jobs are located within a city would therefore facilitate an understanding of the factors that boost local employment. More specifically, knowing where enterprises locate their branches and establishments tells us how many people and firms are reachable within certain time periods, which is important for firms to understand in respect of their potential labour, consumer and supplier markets.

Accurate data also enables cities to assess their economic potential and determine housing and transport policies and interventions in order to connect people to jobs and economic opportunities. Combining job and household locations allows for a cost evaluation of the disconnect between jobs, housing and transport so that potential transport improvements may be accurately assessed (Goswami and Lall, 2016). Since the location of jobs affects a city's land values, data on firm and job location can be used to assess the cost of regulations and policies to ameliorate distortions in urban land allocation (Lall et al., 2017). That said, the COVID-19 pandemic has had a significant influence on remote working, and it is still to be seen how greater automation and 'working from home' will reshape urban space economies, creating the need for 'disruptive innovation' in urban planning.

Data collection constraints

South Africa has limited data on the location of jobs and production at the branch (establishment) level as opposed to the head-office level. Large group conglomerates in finance, construction and mining as well as multi-establishment firms, particularly in retail, banking and insurance, are significant in South Africa's economy. These companies (e.g. Pick n Pay, Shoprite Checkers, Absa and Nedbank) all report the location of jobs and production at the national head-office level rather than at the local establishment level of their branches.

As South Africa's agency for official statistics, StatsSA produces GDP and price (inflation) data for annual national accounting purposes. StatsSA draws samples for its annual economic surveys from its Business Sampling Frame. The Business Sampling Frame, in turn, draws from the SARS value-added tax (VAT) database, as well as from extensive large-group

4 For instance, Sinclair-Smith and Turok (2012) undertook a detailed spatial analysis of urban economic dynamics for Cape Town using anonymised administrative tax data for the period 2001–2005 drawn from the Cape Town database for the RSC levy. The RSC levy's incidence displayed a dominant 'head-office effect' with nearly 60% of total collections favouring municipalities that had a strong head-office bias, such as Johannesburg and Cape Town (FFC, 2013; National Treasury, 2009; Bahl et al., 2003). However, the analysis could not be updated given the underlying data source.

The 2019 draft National Spatial Development Framework draws on the Council for Scientific and Industrial Research's (CSIR) MesoZone 2018v2 database and algorithm, which uses gross geographic value added (GVA) and employment data produced at the local municipal level (by private-sector data provider Quantec) and assigns it to a mesozone-grid layer. The resultant data mapping provides an estimation of employment per sector (excluding construction) at the mesozone level or the potential number of employment opportunities at the place at which people will work (CSIR, 2020).

The CSIR (2020) MesoZone 2018v2 database is also used as the underlying spatial-economic database for the 2016 *State of South African Cities Report*, produced by the South African Cities Network (e.g. SACN, 2016, 2011); and the IHS Global Insight database was used in the spatial-economy background research report for the 2016 Integrated Urban Development Framework (CoGTA, 2016; Turok and Borel-Saladin, 2013).

The Gauteng City-Region Observatory (GCRO) utilised GVA data from the CSIR's Geospatial Analysis Platform (GAP) mesozone database (which in turn draws on modelled Quantec data for economic activity and employment at the municipal level) and from the AfriGIS 2010 Bizcount dataset on the number of businesses to define the economic components of core and peripheral maps for the Gauteng City-Region (Harrison et. al., 2014).

Similarly, Götz and Todes (2014) draw on the CSIR's GAP GVA data and the AfriGIS 2010 BizCount data in their analysis of the changing nature of Johannesburg's urban space economy, highlighting that a key value-add of the BizCount data is its detailed Standard Industrial Classification (SIC)-level categorisation of firm-level activity, which enables sectoral composition and distribution analyses.

and multi-establishment profiling, to determine the coverage and size of enterprises in the country. The StatsSA Business Sampling Frame uses the legal unit (or enterprise unit) as the sampling unit to produce estimates for enterprises at national level. However, this only includes enterprises in the formal sector with a turnover of R1 million or more (the threshold for VAT registration).

The Business Sampling Frame structure provides for data collation and classification at three levels: (1) enterprise units that are registered for VAT; (2) kind of activity units; and (3) geographical units that form a single-activity part of an enterprise operating at a particular geographic location (equivalent to an establishment level). However, it does not have sufficient coverage or completeness of data in its geographical unit frame to enable sampling at the local level (StatsSA, 2003). As a result, StatsSA is constrained to use the enterprise unit as the sampling unit in official economic national survey statistics. The aggregation of reporting in employment and production figures in national economic survey statistics (the 'head-office effect') is then carried through to any modelled sub-national economic data on economic activity and employment that is produced by private-sector data providers such as Quantec and IHS Global Insight. These companies draw on StatsSA's economic survey data as their underlying data source. The same limitation occurs in the Council for Scientific and Industrial Research's MesoZone 2018v2 database (CSIR, 2020), which uses Quantec data for employment and economic activity at the local level.

Consequently, South Africa does not produce an official annual economic survey or census *at the establishment level*, unlike many member countries of the Organisation for Economic Co-operation and Development (OECD) such as the United Kingdom, the United States and Canada. The size and complexity of South Africa's economy combined with the undercoverage of the geographical unit in the Business Sampling Frame also prohibits donor or externally financed establishment-level surveys, which are routinely undertaken for smaller economies such as Ethiopia, Uganda and Tanzania in the World Bank's Enterprise Surveys.

In fact, the recent 2020 Enterprise Survey for South Africa (World Bank, 2021) experienced similar head-office constraints when drawing its sampling frame for the survey. Rather than using official data, the survey had to purchase a sampling frame through a third party. This enabled them to then draw a representative sample for the survey, which interviewed 1 097 firms across Gauteng, KwaZulu-Natal, the Western Cape and the Eastern Cape between December 2019 and February 2021. The survey aimed to assess the challenges facing businesses in South Africa's private sector with respect to productivity, growth, job creation and innovation. Of the firms surveyed, almost 36% were medium-sized (20–99 employees) and 13% were large (100+ employees). Notably, the Enterprise Survey (World Bank, 2021) does not provide detailed city economic data at the establishment level.

This section has detailed the root cause for the paucity of official city economic statistics in South Africa – that is, the undercoverage of data collected at the geographical establishment level in the StatsSA Business Sampling Frame. Appreciating that constrained public finances will not enable the extension of the Business Sampling Frame at the geographical level in the medium term, urbanists have turned towards alternative and innovative approaches to obtain the intra-city economic data that they need.

Accurate data enables cities to assess their economic potential and determine housing and transport policies and interventions

Anonymised and geocoded data

As the alternative to general survey data, the next viable step in making city economic data accessible is to use anonymised and geocoded economic data collected during public administrative routines. In 2017, the first United Nations World Data Forum highlighted the global shift towards the use of anonymised administrative data sources combined with big data (particularly mobile phone and banking data) and sensor data (such as road traffic and water-level data) for statistical analysis – made possible by significant technological advances in automation and data management.

Using anonymised administrative data in statistical analysis not only improves the integrity and efficiency of the statistical production process, it also reduces the costs of statistical collection. In addition, it minimises the burden placed on respondents, particularly on businesses in firm-level surveys (Malmidin, 2017; Ruotsalainen, 2017). By its nature, administrative data has a longitudinal structure that allows informative time-series analysis. In addition, given the high rates of non-response and under-reporting in survey data collection, administrative data also offers more credible and accurate information than survey sources (Malmidin, 2017; Pieterse et al., 2016).

South Africa has some experience in using anonymised administrative data since tax administrative data has been used to determine the StatsSA Business Sampling Frame for economic surveys for over a decade. Starting in the early 2000s, SARS administrative systems were overhauled and modernised. As highlighted by Gavin et al. (2013), this has significantly improved the quality of information that can be extracted from the tax system. This improved data is also more accessible for analysis, in turn enabling SARS to publish anonymised and aggregated tax statistics at municipal level in the annual tax statistics bulletins, informing a degree of sub-national economic analysis (Gavin et al., 2013).

As a new source of microdata for research and analysis purposes, the release of anonymised tax statistics has generated both significant interest and utilisation among policy economists and academic researchers. In 2015, the National Treasury and SARS collaborated with the United Nations University–World Institute for Development Economics Research (UNU-WIDER), under the Regional Growth and Development in Southern Africa programme, to establish a panel tax dataset, created by merging four tax data sources – company income-tax (CIT) data, employee pay-as-you earn (PAYE) tax certificate data, VAT data and customs records.

Jumping ahead in our narrative, continued efforts to develop South Africa's administrative micro-level taxation data through further collaboration with UNU-WIDER, under the Southern Africa–Towards Inclusive Economic Development (SA-TIED) programme over the period 2017–2020, have made administrative data available to researchers through the National Treasury Secure Data Facility (NT-SDF).[5] In 2017, the National Treasury and SARS also partnered to geocode the PAYE dataset, enabling its utilisation for more granular spatial-economic research and analyses.

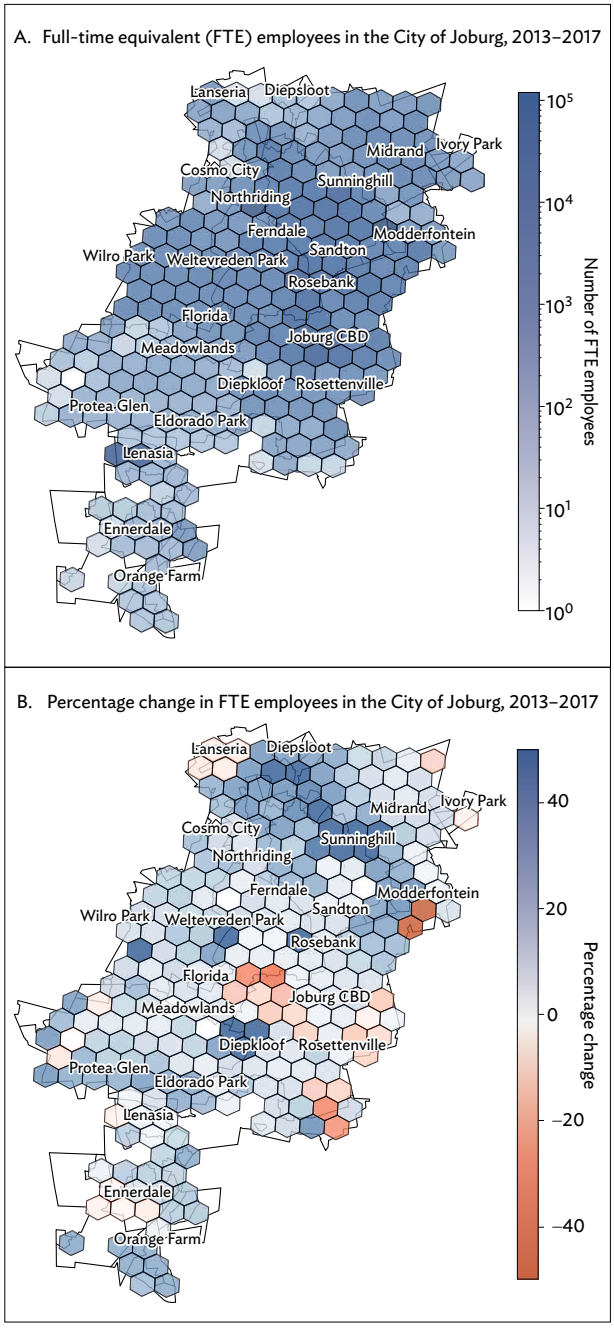
The recent launch of the *2021 City Spatialised Economic Data Reports* (CSP, 2021) is a foundational milestone in this respect since they make available processed anonymised and geocoded tax data and city profiles. It was possible to produce the *2021 City Spatialised Economic Data Reports* because of the efforts of a committed group led by the National Treasury's CSP and the broader ERLN, who were part of the city data journey to date and whose efforts are described in the following sections. A significant amount of person-energy and time has been required to access and align public administrative datasets and make them available for public research purposes.

Advances in data technology do not mean that a 'data genie' suddenly appears, unlocking access to aligned, cleaned and interoperable administrative

5 The anonymised datasets available in the NT-SDF include: (1) the SARS-NT/CIT-IRP5 panel for the period 2008–2016; (2) the individual panel (IRP5 and ITR12) for the period 2011–2018; (3) CIT for the period 2008–2016; (4) VAT for the period 2009–2017; (5) customs data for the period 2009–2017; and (5) PAYE (payroll or IRP5) data for the period 2009–2017 (SA-TIED, 2020).

Figure 3: Aggregated and spatialised administrative tax data

SOURCE: *Metro-level report: City of Joburg*, p. 8 (CSP, 2021); redrawn by Lumina Datamatics



data. Rather, the preparation of such datasets is complex and messy. In reality, administrative datasets are contained in different IT systems across government and the wider public sector, many of which are outdated and dysfunctional as well as overseen by multiple officials. In addition, the datasets within these systems are collected and maintained for different bureaucratic purposes. Finally, given the personal identity information contained within the datasets, they are often shrouded due to privacy concerns and protected by legislation such as the Protection of Personal Information Act (No. 4 of 2013).

Administrative datasets are therefore the outcome of a relational process, aptly described by Boyd (2022) as 'data-as-assemblage'. When one takes time to watch, listen to and understand the actions of the people and processes that outline and use ('own') administrative datasets, the 'hidden obvious' emerges into view, allowing insight into how administrative data can be collated and utilised for research purposes. Contemporary ethnographic studies help to reveal the assumptions and daily practices underlying data management processes, services and products, and enable the creative insight to design potential solutions to persistent challenges. These studies show that when people do their jobs, or any routine activity, much of the 'doing' becomes invisible and almost rote in nature. Notably, these officials are often working around a particular problem, and it is precisely this gap in knowledge that becomes an opportunity for innovation (Isaacs, 2013).

The decade-long adventure in city economic data (continued in the following sections of this paper) elucidates the potential of using official administrative data for city spatial-economic analysis as an alternative to survey data. However, the narrative also underscores the extent of the challenge. It highlights the effort required, first, to assemble different data custodians and users in government and the wider public sector, and second, to encourage collaboration that aligns complementary administrative datasets in a way which, while protecting identity confidentiality, opens up the wealth of public-sector data to urban research and analysis.

City government advocacy

South Africa's city governments and urbanists have long advocated for better access to city economic data. Numerous conversations have lamented the lack of reliable, publicly available economic-activity data at the intra-city level.

City governments such as the metros of eThekwini, Cape Town and Johannesburg are strengthening their analytical capabilities to deepen spatial-economic research and inform robust urban economic policy development. In this respect, credible evidence-based spatial-economic analysis at the intra-city, main-place level enables city governments to analyse the performance and potential of economic nodes across city space, ensuring intergovernmental coordination in the planning and implementation of urban investments.

Accordingly, eThekwini and Cape Town have both launched open data portals that include spatial datasets related to building plans, valuation rolls and business licensing, to name a few. However, spatialised data on economic activity and employment is not available on these portals. Instead, eThekwini, Cape Town and Johannesburg have all drawn on modelled data at the municipal level (supplied by private-sector operations IHS Global Insight and Quantec) to produce their monthly and quarterly economic reviews and updates.[6]

6 *eThekwini Economy at a Glance* is available on a monthly basis (e.g. see eThekwini Municipality, 2021). The Economic Performance Indicators for Cape Town (EPIC) are produced on a quarterly basis (e.g. see City of Cape Town, 2021). The City of Johannesburg produces monthly and quarterly economic reviews for internal city economic planning purposes only.

The insistent clamour for improved access to city economic data was heightened in 2018 when the National Treasury's CSP introduced a requirement in its *Guidance for the 2017 and 2018 City Built Environment Performance Plans* (BEPPs)[7] to incorporate disaggregated economic data on economic nodes at a sub-national level and to map these nodes against integration zones in order to analyse city economic performance and potential within the broader city space economy (National Treasury, 2018a).

The CSP was aware at the time of issuing the 2017 and 2018 BEPP guidelines that disaggregated economic data at the sub-national level was not easily available in official survey statistics from StatsSA, and that modelled data could be purchased from private-sector data providers. However, the requirement for disaggregated city economic data was inserted in the guidelines to galvanise city and intergovernmental efforts to access city economic data through their collaborations within the ERLN's Technical Working Group on Data.

The next two sections recount the CSP and ERLN's collaborative efforts to find a solution to obtaining city spatial economic data for the World Bank's Urbanisation Review of South Africa. What seemed like a reasonable request at the outset turned out to be a herculean task. The CSP and ERLN confronted governance challenges, fragmentation across government systems and silos, capacity constraints, and often a lack of political will itself to understand and collaborate on solutions that address the city economic data conundrum.

Photograph by Beth Pulane Crankshaw

7 City BEPPs were first introduced in the 2011/12 financial year as an eligibility requirement in respect of the Urban Settlements Development Grant in the Division of Revenue Act (No. 5 of 2012). The BEPPs enable cities to align planning for the various built environment financial grants within the municipal space, and to enhance intergovernmental coordination in the planning and implementation of urban investments. In 2014, guidelines were introduced to assist cities in the development of their BEPPs, addressing key weaknesses in planning and budgeting frameworks. The 2017 guidelines emphasise the BEPPs as a key instrument in urban spatial transformation, introducing priority integration zones as the focus of public-sector investment. The 2017 guidelines further require the city BEPPs to present analyses of the performance and potential of metro economic nodes within integration zones and city space more broadly, and point to the ERLN's Technical Working Group on Data for analytical tools and support in this regard.

Photograph by Siarhei / Adobe Stock

StatsSA and the Urbanisation Review of South Africa

StatsSA and the Urbanisation Review of South Africa

Structured intergovernmental collaboration

In late 2014, the ERLN convened two inception workshops to enhance metropolitan government access to national government administrative data. Thereafter, the workshops were held quarterly, through to 2017. These workshops, facilitated by GTAC, brought together participants from metropolitan municipalities, national government departments, the government agencies StatsSA and SARS, and academic and state research institutions including the GCRO, the University of Cape Town's Data First, the Human Sciences Research Council (HSRC) and the CSIR.

The initial 2014 workshops highlighted the need for structured intergovernmental collaboration regarding metropolitan and regional economic data and analysis. The following issues meriting joint attention were subsequently outlined in a memorandum submitted to the Budget Forum[8] for endorsement:

- *Improving sub-metropolitan economic information and analysis through accessing anonymised taxation and UIF data.*
  The most critical gap in urban information is sub-metropolitan information on economic activity and employment. In this regard, it was noted that there had been a significant loss of sub-metropolitan firm-level data since the phasing out of the RSC levy in 2006 (see fn 3). It was further accepted that the tax data collected by SARS is currently the best potential source of firm-level data and that this data could be aggregated to various sub-metropolitan scales while preserving the necessary confidentiality. The tax data, however, is limited in that information on larger companies is generally reported at head-office level and location despite enterprises having multiple establishments and thus points of employment and economic activity distributed across the country. However, the working group thought that this constraint could be addressed by linking the tax data with other data sources such as UIF data.
- *Ensuring that national economic surveys provide statistically meaningful metropolitan and sub-metropolitan information.*
  It was pointed out that the samples used by StatsSA in their various surveys were often not large enough to provide accurate metropolitan or sub-metropolitan information, and that significant improvements in the information required to track economic changes between the censuses could be achieved with larger sample sizes.
- *Legislative change to remove unnecessary regulatory obstacles to the sharing of potentially useful administrative data.*
  SARS and StatsSA had been working on amending legislation to try to ameliorate this situation. It was agreed that input from the metropolitan municipalities would be useful, and that SARS and StatsSA should be requested to engage with these government agencies in this legislative process.

8 The Budget Forum is a platform that comprises the Minister of Finance, the nine provincial Members of the Executive Council (MECs) for Finance and representatives of the South African Local Government Association (SALGA).

- *Mining and collating municipal data for purposes of socio-economic analysis.*
  The metropolitan municipalities collect substantial amounts of administrative data themselves, including electricity, water, property rates and building-plan information. However, extensive data cleaning and analysis is required before useful socio-economic information can be generated. This is best done by the metropolitan municipalities with support from specialist agencies such as StatsSA.
- *Public-sector data as an economic tool.*
  It was noted that making relevant governmental data available to the public had the potential to create direct economic value by providing firms with up-to-date information that could be used to develop new products and services. This move would also enhance national and regional economic competitiveness by reducing the risk and cost of doing business. However, clear guidelines on accessing the available data were needed.

Urbanisation Review of South Africa 2016–2018

The deliberations and efforts recounted so far reached a peak during the World Bank Urbanisation Review Inception Mission to South Africa in April 2016. In preparation for its collaboration with the World Bank to undertake the Urbanisation Review, the National Treasury made a further call for disaggregated city economic data. The Urbanisation Review's meetings and workshops drew on the evidence-based analysis provided by the South African team, first, to quantify the economic, social and fiscal costs of the current patterns of the urban space economy, and second, to identify the chief institutional and fiscal challenges in housing, transport and industrial development strategies (World Bank, 2016a).

While the Inception Mission was impressed by the overall quality of technical skills, data and analytics in the country, they noted that the lack of publicly available economic data sources for determining the locations of jobs and business was likely to undermine the quality of analysis needed for urban policy-making (World Bank, 2016a). The mission team examined a variety of data sources that report jobs and economic activity, indicating where head offices and holding companies are registered but not where individual establishments are located throughout the country. This information constraint hampered the scope of the Urbanisation Review analyses meant to examine the productivity differentials and barriers to growth across the urban space economy. This outstanding data gap was specified as

> obtaining data on the detailed spatial location of the average employment in each metropolitan area by main place level, classified by sector of employment (ISIC, Revision 4 – 4-digit preferable). Addition of information on employment by skill level and wages by skill level would be advantageous, as would any information on inputs/outputs and prices, and potentially sales. (World Bank, 2016b)

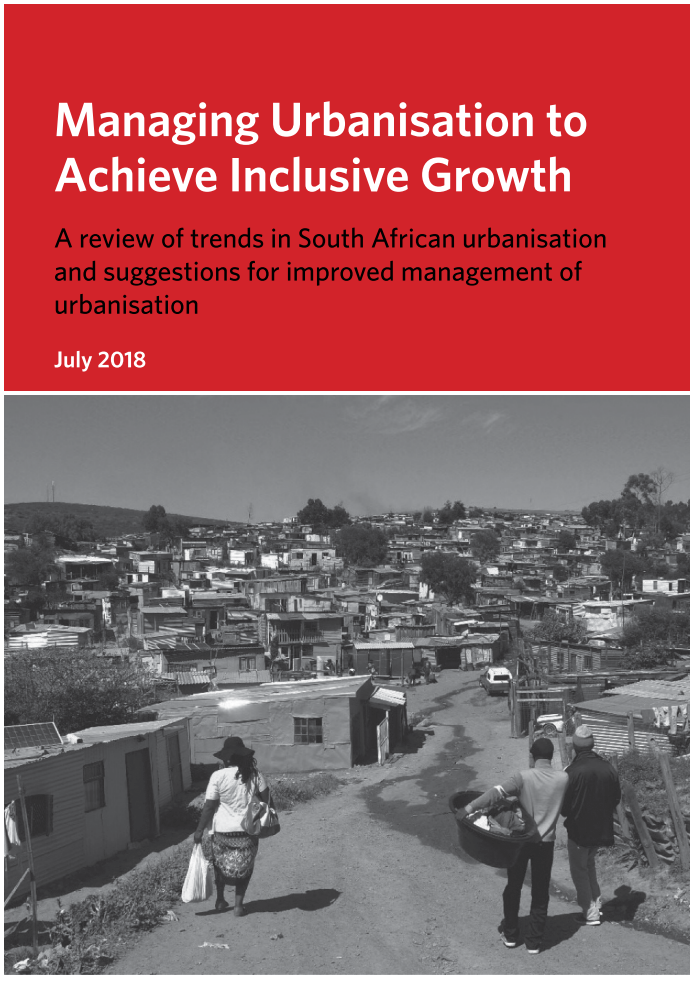
This important issue requires some explanation. The International Standard Industrial Classification of All Economic Activities (ISIC) Revision 4 is the international reference classification of productive activities. Its main purpose is to provide a set of activity categories that can be utilised for collecting and reporting economic statistics according to these categories. The majority of countries worldwide use the ISIC as their national activity classification or have developed national classifications derived from the ISIC (UN DESA, 2008). Economic activities

Extensive data cleaning and analysis is required before useful socio-economic information can be generated

are subdivided in a hierarchical, four-level structure of mutually exclusive categories to facilitate data collection, presentation and analysis at detailed levels of economic activity in an internationally comparable and standardised way. The categories at the highest level are called sections, for example agriculture, forestry and fishing is section A, and manufacturing, section C. The classification is then organised into successively more detailed categories that are numerically coded: two-digit divisions; three-digit groups; and at the greatest level of detail, four-digit classes. The economic classification is used to further classify statistical units such as enterprise units or establishments (geographical units) according to the type of economic activity in which they mainly engage.

The importance of ISIC classification for the World Bank Urbanisation Review added further weight to increasing city insistence on the need for their access to official sources of city data.

Figure 4: Various reports comprising the World Bank Urbanisation Review of South Africa were published in 2018
SOURCE: National Treasury (2018b)



This realisation spurred on the CSP and the ERLN Technical Working Group on Data in their efforts in this regard, ahead of the then imminent finalisation and publication of the Urbanisation Review in 2018 (see the executive summary of the Review, National Treasury, 2018b).

Involving StatsSA in the Urbanisation Review

From 2014 to 2017, the workshops convened by the ERLN Technical Working Group on Data continued on a quarterly basis. At the July 2016 meeting, with fresh impetus from the World Bank's involvement, the discussions focused on finding a workable solution to obtaining the disaggregated spatial-economic data required. StatsSA raised the possibility that they could undertake a user-pays establishment-level survey on behalf of the CSP for the Urbanisation Review. And later that month, a video conference – with the World Bank Urbanisation Review team, the University of Oxford, StatsSA, SARS, the National Treasury's CSP, the Technical Working Group on Data and a representative of the University of KwaZulu-Natal – was held to determine the Urbanisation Review's spatial-economic data requirements and to assist the StatsSA team to prepare a response as to how they would approach undertaking their proposed user-pays establishment-level survey.

Search for a sampling frame

All parties agreed that a robust sampling frame for the survey needed a first-level census of establishments by metropolitan area. Such a survey would require fieldworker entry to obtain the business/establishment name and main business activity (classified by SIC 7, StatsSA's version of ISIC 4[9]). Each establishment in this first-level census would need to be classified by geocode, business name and main business activity (i.e. SIC 7). Then, a second-level sample frame would need to be drawn from the full establishment census by metropolitan area.

In order to generate a sampling frame of firms for the survey, StatsSA initially investigated whether they would be able, first, to utilise geocoded VAT data to extract the eight metro areas by their postal code, and second, to remove areas that are residential and agricultural using mapped census data. The approach to provide the centres of employment per metro visually would draw on population density information (where people live) available at a small area level from the census data, and against such, map how many people per residential area are employed. This method could then provide an overall geographic map of where people work versus where they live.

To accomplish this land-use approach, StatsSA proposed drawing on its dwelling frame. However, while the dwelling frame does provide an indication of which locations are involved in business, it does not indicate either the type of economic activity or the name of the business. Furthermore, at the time, StatsSA had not tested whether the frame accurately represented all the businesses in an area. The StatsSA dwelling frame uses both primary as well as secondary data sources (such as municipal property-valuation rolls) to infer economic activity at a location point through an unpacking process requiring fieldworker entry or an aerial photograph of the area. However, if there are multiple businesses located within the same structure, the dwelling frame only captures general business activity and does not capture the types, names or sectors of the businesses. In effect, this unpacking process only covers less than a tenth of the total dwelling frame.

9 StatsSA's Standard Classification of All Economic Activity, 7th edition (SIC 7) (StatsSA, 2012) is based on the latest revision of ISIC Revision 4, released in 2008. The basic coding of the South African classification system has been changed to align it with ISIC 4. The classification still goes up to a five-digit level, but the categories have been changed to section, division, group, class and sub-class.

Additional sources of establishment-level information

If followed, this land-use approach to the sampling frame would have required StatsSA to consider other sources of information to supplement the dwelling frame, such as from OpenStreetMap. However, this supplementary approach would only have provided a business name – and not necessarily the sector in which it was active.

Further queries raised concerns as to whether land-use cover would be able to build the establishment-level sampling frame in the way StatsSA suggested. The verification of sectoral use is complex and, as we have seen, not predicted well by land-use category. More specifically, multi-occupancy structures and business premises that do not externally appear as business premises make the physical verification of business activity extremely challenging. If a sample of firms by sector and size is required, and not just by spatial area, it is important to have a method for attaching attributes to building structures in the dwelling frame. Whereas the StatsSA dwelling frame does this in terms of the population census, there is no equivalent for collecting business data.

For this reason, a census was required that would obtain the basic attributes allocated to an address/structure. A representative sample could then be drawn to get more data in a second step, analogous to the way StatsSA undertakes the Survey of Employers and the Self-Employed (SESE). The SESE is an enterprise-based survey that collects data on micro and small businesses in South Africa, covering information about businesses in the informal sector to gain an understanding of their operations and access to services. The SESE focuses on small and micro businesses that are not registered for VAT and are therefore excluded from StatsSA's Business

Figure 5: Informal businesses and many small and micro businesses are not registered for VAT
Photograph by Kgoa Mashego

Sampling Frame. The SESE first identifies non-VAT registered businesses through a household survey and then, for the actual survey, a sample is drawn from the businesses identified in the first round (StatsSA, 2017).

In light of these challenges, stakeholders then discussed whether it would be possible to draw on the StatsSA dwelling frame as a source and, together with the Community Survey 2016, undertake a census of all dwelling units to develop a register of businesses by geographic area for a specific metro. This census could identify (1) the geocoded location of the business; (2) the business type; (3) the company registration number if registered at the Companies and Intellectual Property Commission; and (4) the business's sector and SIC 7 economic classifications. This listing could then be used to extend StatsSA's business register for each metro.

However, disagreement arose between the stakeholders as to the next step. Could a randomised sampling approach be used to develop a sample frame for further detailed interviews that would be statistically representative of firms in the metro area? Or would the frame have to be constructed by walking the area and ensuring physical fieldworker entry into a business to obtain a first level of information? The latter approach was considered important for the frame's completeness and statistical accuracy.

The merits of the two survey levels would be important to establish since a branch within a multi-establishment enterprise might not have information on other branches. The proposed two-step survey would need to undertake interviews with businesses in the sampling frame to collate data on (1) the size of business; (2) the establishment-level detail of the type of operations for the SIC coding if a multi-establishment enterprise; (3) the average employment during the last calendar or financial year; and on (4) the average salary/wage bill at the establishment over the last calendar or financial year.

Lessons from a 2014 survey

This two-step approach had already been used in a survey on 26 industrial nodes in Johannesburg undertaken by the Centre for Competition, Regulation and Economic Development (CCRED) for the City of Johannesburg in 2014 (Kaziboni et al., 2015). More specifically, the CCRED study had undertaken an establishment census of all industrial nodes, followed by a survey of a representative sample of manufacturing and related services. The survey team had then conducted a series of in-depth interviews in seven nodes to develop detailed economic profiles of each node.

CCRED's first-step census was undertaken through a street-by-street field investigation of all business activity in 26 industrial nodes selected by the City of Johannesburg over a five-week period in August and September 2015. The exercise was conducted by a team of specialist fieldworkers, who identified and described all businesses in the nodes and captured firms' contact details. Additional steps were taken to verify the data's accuracy by cross-checking information against available business cards and double-checking categorisations that appeared incorrect or where data was missing through desktop research. Follow-up telephone calls were made to confirm firms' information or to obtain missing data.

CCRED's second-step manufacturing survey had 45 questions whereas the manufacturing-related services questionnaire listed 40. Both surveys were administered in English (only) using an online platform, SurveyMonkey.

As input into the design of a StatsSA user-pays firm-level survey for the Urbanisation Review, CCRED provided both their first-step census scoping questionnaire and the second-step manufacturing survey and manufacturing-related services questionnaires to the National Treasury's CSP and the ERLN's Technical Working Group on Data team.

Businesses that are not registered for VAT are excluded from StatsSA's Business Sampling Frame

Survey coverage, cost and budget

Stakeholders then queried whether StatsSA's proposed city user-pays survey should cover both informal (unregistered) and formal businesses. If only covering formal businesses, it would be important to include businesses that are co-located within dwellings as well as businesses located at 'commercial' or 'industrial' premises, depending on the size of business to be covered in the survey (a lower limit of five employees could potentially exclude the self-employed). As developing the census to create the frame would be costly, the sample size was a further cost driver in the overall exercise. In addition, the questionnaire would need to be piloted prior to the actual survey.

These challenges notwithstanding, the terms-of-reference for the user-pays establishment-level survey were submitted to StatsSA in mid-August 2016, and StatsSA was asked to respond to a proposed budget for undertaking the survey.

Confronting the 'head-office effect'

The National Treasury and ERLN then met with the StatsSA Business Register team to understand the way in which the Business Sampling Frame is constructed from VAT administrative data and large-company profiling. From the SARS perspective, multi-establishment companies can either register a single VAT number for the head office covering all the establishments within the enterprise, or they can register one VAT number for head office and additional numbers for each of its local branches. Most companies, however, only register a single VAT number for the head office – hence the 'head-office effect' in StatsSA's official economic surveys, which use the enterprise (head-office) unit rather than the geographical establishment units for survey sampling. As has already been noted, this 'head-office effect' is then carried over into any modelled sub-national economic datasets produced by private-sector data providers such as Quantec and IHS Global Insight.

The National Treasury and ERLN also engaged with StatsSA to better understand the institution's approach to profiling enterprise groups. While these activities do indeed collect a large percentage of establishment-level addresses, turnover, employment statistics and SIC 7 classifications within the geographic level of the Business Sampling Frame, the data obtained is not sufficiently complete for using the geographical level as a sampling unit – a critical requirement for the Urbanisation Review.

StatsSA bows out

Further discussions were held on the approach StatsSA could follow to develop the enterprise census as the first step. However, by the end of 2016, it was generally agreed that StatsSA did not have sufficient capacity to undertake the work envisaged in the terms-of-reference. More specifically, StatsSA lacked the capacity to undertake the required full census of establishments within a metropolitan area given that the undercoverage of the geographical level in the Business Sampling Frame limited its potential to be used as a sampling unit. The 'data pathfinders' were thus obliged to pursue other strategies for obtaining the city economic data required for the Urbanisation Review.

Photograph by Clive Hassall

Pretoria street map by Hairem / Shutterstock

Continued collaboration with SARS

Continued collaboration with SARS

SARS anonymised and geocoded tax data

During 2017, after the decision to abandon the StatsSA plan to conduct a user-pays establishment-level survey, the CSP and ERLN focused on collaborating with SARS to draw on anonymised and geocoded economic administrative data (collected in the course of public administration) to obtain information on the detailed spatial location of firms and jobs for the World Bank Urbanisation Review.

The CSP and ERLN collaboration with SARS was a natural progression from the steps that SARS had already taken to use geocoded and anonymised tax data for research purposes. At the first ERLN-convened workshop in 2014, the SARS team had pointed out the potential of using the geographic addresses contained in their tax data to identify the location of jobs. SARS had confirmed at the time that as part of its anonymisation interventions, it was exploring the geocoding of address fields in the tax data to include spatial elements and was interested in enabling the wider use of anonymised tax records for policy-relevant research in line with tax administration practice worldwide (e.g. in the United Kingdom and New Zealand).

These earlier efforts had led, in 2015, to SARS and the National Treasury's collaboration with UNU-WIDER to establish a panel tax dataset: the SARS-National Treasury Panel (hereafter, the SARS-NT Panel). The SARS-NT Panel was created by merging four tax data sources: (1) CIT data; (2) employee PAYE tax certificate data; (3) VAT data; and (4) customs records. Initially, the anonymised firm- and PAYE-level data was only accessed through a single server and two computer terminals at the National Treasury that UNU-WIDER had donated for the purposes of research under the three-year collaboration programme.

SARS led two key undertakings at the Technical Working Group on Data from 2015: the Economic Classification of Businesses and the Use of Physical Addresses to Disaggregate Economic Statistics projects. During the first project, SARS began the process of migrating from its former SIC 5 classifications to the internationally compatible SIC 7 standard (ISIC 4, see fn 9) in order to provide standardised and accurate business classifications of tax-registered businesses. After transitioning to SIC 7, StatsSA, as the owner of the standard economic classification, would then work with the metros to update the economic classifications of businesses in their geographic areas.

In this respect, SARS was keen to explore how national and city administrative data could be combined in research approaches. For instance, aligning national and metro data via the economic

Geographically disaggregated economic activity could be modelled using metro-held information and national data

classification of businesses would significantly improve the quality of data used in decision-making. Geographically disaggregated economic activity could be modelled using metro-held information (e.g. where commercial buildings are planned/completed) and national data (e.g. the estimated VAT from tax data).

While the metros did not collaborate as actively with SARS as originally hoped, SARS was able to complete most of the required work on the economic classification of data. Lessons learned included that (1) the self-reporting of economic classification was not accurate, and (2) relatively short training interventions (undertaken by StatsSA) were needed to equip staff to significantly improve the accuracy of economic classifications.

The second undertaking, the Use of Physical Addresses to Disaggregate Economic Statistics project, focused on using appropriate methods to geocode semi-structured (poor quality) address data, and develop recommendations for improving data quality to enable automatic geocoding over time. To undertake spatial socio-economic analysis at the sub-municipal level, SARS aimed to use physical address data for individuals, households and businesses.

As the owner of the physical address data of registered tax payers (individuals and businesses), SARS was interested in collaborating with metros to improve the quality of SARS address data as well as to develop methodologies that could match residential addresses to the geography of economic activity. SARS had engaged with StatsSA in this respect but had not been able to confirm an agreed approach to innovative geocoding methodologies. Possibilities included (1) using the locations of new CIT, VAT and PAYE registrations as a proxy for where new economic activity is being initiated and where employment by PAYE-registered employers is taking place; and (2) strengthening initiatives to geocode PAYE tax certificates according to the physical address of the recipient.

At the Technical Working Group on Data workshops during the World Bank Urbanisation Review mission to South Africa in early 2016, SARS presented their efforts to match residential addresses captured in the administration of PAYE to the 'national address database'. They raised their concern that a high percentage (10–40%) were not matched because of the different databases used. Accordingly, SARS agreed that their matching geocoding algorithms needed to be improved considerably to ensure greater accuracy, and that their algorithms were not scalable to large databases in their current form. In particular, SARS raised the challenges of using postal codes to obtain the municipal/main-place of the taxpayer.[10]

In this respect, SARS presented lessons learned from previous attempts to use postal codes to geocode VAT registration data at the municipal level, noting the challenges in the accuracy of postal code data that clearly link a particular postal code to a specific main-place within a municipality. SARS raised as a concern the lack of structure in the way tax payers complete address fields, which necessitates using a sophisticated automated geocoding algorithm for tax address fields other than the postal code (ERLN, 2015).

10 Geocoding accuracy depends on the credibility of the address and other point data on which it is built. Address data often comes in different styles and formats, such as street intersections, house numbers with street names, units within a business park along a specified street name, or postal codes.

South Africa has a national address standard defined by SANS 1883, published by the South African Bureau of Standards. SANS 1883 defines the data elements and formats of addresses, and provides guidelines for address allocation and maintenance.

However, while the SARS Business Requirements Specification (BRS) for PAYE Employer Reconciliation includes fields for employee physical work address details, of the fields capturing these details – unit number (code: 3144), complex (3145), street number (3146), suburb/district (3148) and city/town (3149) – only the street name (3147) and postal code (3150) are mandatory (SARS, 2021). Also, it is unclear whether the BRS specifies the address structure in line with the SANS 1883 requirement.

While SARS does have a geocoding algorithm, it is not integrated into the SARS e-filing process, so address fields are manually entered by employers completing their employer reconciliations. Manually captured address data that is not forced to comply to specific address structure standards (notably SAN 1883) often end up as incorrect ('dirty') data, with the incorrect spelling of street or suburb names, the street number after rather than before the street name, to name a few problems.

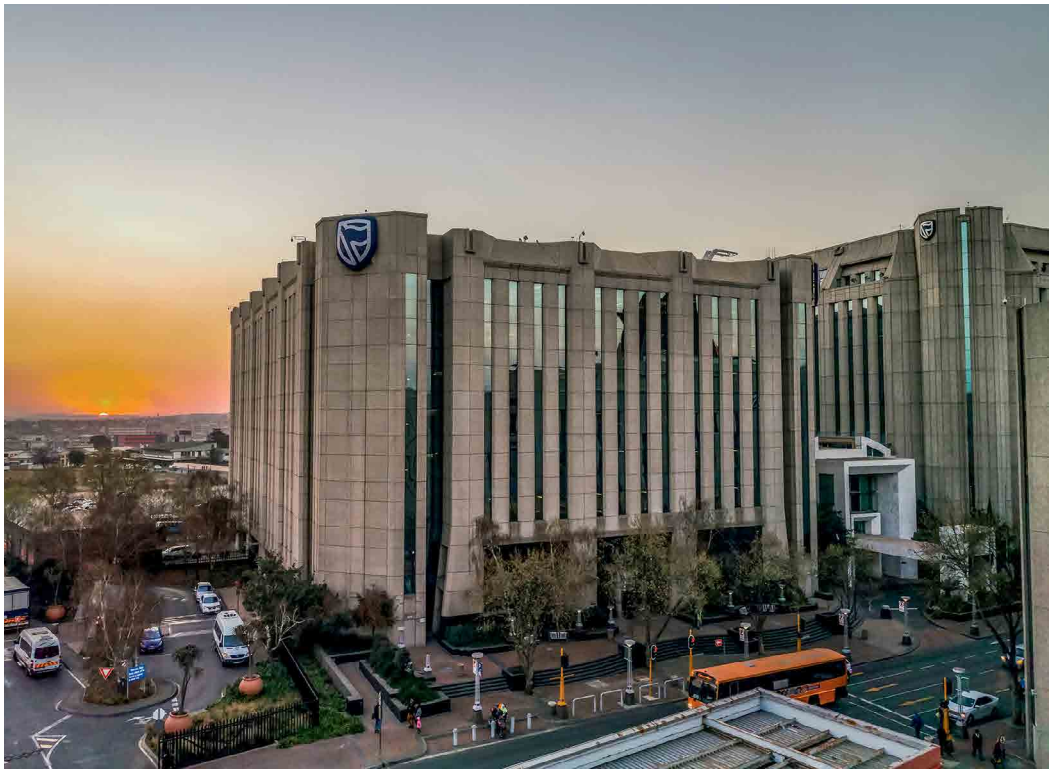
Geocoding SARS PAYE tax data

SARS and the National Treasury's continued collaboration on the potential of using geographically disaggregated tax data fed into presentations to the Technical Working Group on Data that reviewed the types of geographic information contained in the tax data. It was noted that tax payers (businesses, individuals and PAYE-registered employers) provide SARS with both physical and postal addresses when they register for a particular tax as well as on their tax returns. The forms used provide for semi-structured addresses that are aligned to the national address standard.

SARS confirmed that for a CIT registration and return, a single address is provided (typically for the head office) that gives no indication of the geographical extent of the enterprise's business activity. Whereas individuals filing their annual personal income-tax (PIT) returns provide their physical address, not all economically active (or even employed) individuals are required to file a tax return. For VAT, a single address may be provided for each separate branch registration. However, most businesses only have a single registration, and only those businesses with a turnover of R1 million a year are required to register for VAT.

For PAYE, although employers can provide an address for each branch registration, most only register a single branch. PAYE paid over during the tax year is reconciled against the remuneration to, and PAYE withheld from, each individual

Figure 6: Multi-establishment firms report the location of jobs and production at the national head-office level
Photograph by Shams Faraz Amir



employee. This information is captured on individual employees' tax certificates (IRP5s/IT3As), which provide the physical addresses of individual employees and, if different, their postal addresses as well. However, although mandatory, less than half of the addresses geocode 'easily' because the data may be poorly structured. And while the IRP5/IT3A forms do require the physical address of the workplace of the employee, the address fields are poorly populated. From the sample of 2015 large-employer tax certificates, the address-field population rate was only 10%.

In light of these challenges, SARS then pointed to the possibility of tax certificate records supplying a proxy of where jobs are located given that the nation's ten largest employers provide approximately one million out of 13 million jobs across the tax year, and the largest 50 employers provide approximately 2.5 million jobs. However, it is also important to note that about 60% of PAYE-registered employers have only ten or fewer employees (ERLN, 2016).

SARS also presented the results of their initial analysis on using tax certificate records to locate the spatial distribution of jobs in South Africa. While this presentation highlighted the limitations of using address data, SARS had since transferred their focus to other areas. This lack of resolution meant that understanding the different types of address data remained confusing to employers. As a result, many employers continued to use the business address fields to record the work place address of employees; and address data were poorly captured with, for instance, postal codes still appearing in the street name field. To solve this problem, SARS proposed using a geocoding algorithm and spatial modelling to enable the best use of legacy address data. Future effort would need to be invested into ensuring improved compliance, with employers accurately completing the mandatory address fields.

SARS noted the benefits of developing a method that would yield a greater than 90% accuracy rate on the geocoding of poorly structured address data. Such a method should be scalable, allowing the geocoding of files containing tens of millions of addresses. While the Technical Working Group on Data acknowledged SARS's efforts to do this, a level of improvement was still required. SARS further suggested that any method that was developed should then be made available for use within the public sector as a whole and/or distributed under copyright for use in the public domain.

In addition, SARS noted that while the linking of a point location to a physical address (or place of work) was the ideal, linking to suburb/village level (or, worst case, to municipal level) could be considered in cases where a physical address link was not possible. In terms of the location of a job, methods utilising more than one address (e.g. postal and residential and/or workplace) could be explored.

Attempts to procure external assistance

During these presentations, all stakeholders confirmed that the ultimate goal was to validate and structure address data at the point of input. However, there were concerns that the SARS approach thus far was not sufficiently sophisticated to obtain the extent of geocoding required for the Urbanisation Review. Consequently, the National Treasury and SARS agreed that external technical assistance should be sought.[11] Given funding constraints within SARS, the National Treasury and SARS collaborated to procure external technical assistance for geocoding the tax data (in particular, IRP5 data). To be completed on-site at SARS, the exercise would have entailed the geocoding of PAYE tax data legacy addresses, involving the automatic geocoding of the PAYE database. The datasets, approach and algorithm were to have been reviewed after each of four iterations. Sub-place, main-place and ward links were to have been included in each address.

11 Engagements with SARS highlighted concerns around both the strength of the SARS geocoding algorithm and the extent of the underlying 'dirty data' challenge.

The SARS team was hopeful that the 2016 PAYE guidance to employers – requiring greater compliance in respect of work place addresses – would lead to more accurate records for the 2016 tax year. With a reasonable level of compliance, the physical location of an employee's workplace could be ascertained. However, since physical addresses are often poorly structured, they would need to be geocoded using an application programming interface (API)[12] linked to the geocoding algorithm. The API would ensure that employers are not able to electronically submit their forms until they have completed the physical location of the employee's workplace, which the geocoding algorithm would then match to the correct location.

The collaboration between SARS and the CSP required the signing of a memorandum of understanding (MOU) that confirmed their agreement to draw on geocoded and anonymised employer and employee PAYE tax records. This would have entailed determining the detailed spatial location of average employment in each metropolitan area by main-place level, classified by sector of employment. Unfortunately, delays within the National Treasury and SARS meant that the final MOU was only approved in early 2017, making the procurement of external technical assistance to geocode the tax data too late for the Urbanisation Review.

Limitations of the SARS geocoding algorithm

SARS then undertook further internal engagements to determine whether the geocoding service provided by the SARS data analytics team provided a sufficiently robust geocoding outcome for the Urbanisation Review, thereby averting the need for an external service provider to geocode the tax data. These delays emphasised the difficulties of collaborating with SARS at that time,[13] as surfaced in the subsequent Commission of Inquiry into Tax Administration and Governance (Nugent, 2018). Despite these challenges, the CSP continued to follow up with SARS, which confirmed it would provide the geocoded physical work addresses for PAYE tax certificates at the end of 2017.

The SARS geocoding algorithm aimed to provide output at a suburb, main-place, municipality and province level. Only less than 5% of the records seemed to be problematic, the result of data unavailability. The geocoding output was classified according to six levels of accuracy: street, number, suburb, crossing, proximity and 'failed'. SARS quality assurance of the geocoded physical residential address was said to have provided guidance on the changes required in the SARS geocoding algorithm. After reviewing the SARS geocoding outputs, however, the CSP remained concerned that the SARS geocoding algorithm was still only able to provide 65% of the records accurately geocoded to the level required. The SARS data analytics team pointed to the poor quality of address data submitted at source and disputed whether confidence in address data accuracy would increase if the same exercise were to be repeated.

The next step would have been to request the involvement of the SARS Systems and Stakeholder Engagement divisions in a steering committee for the project. However, given escalating governance issues at SARS at the time, the National Treasury decided that it would be best to delay further engagements until there was greater stability at SARS. The absence of city administrative data meant that the World Bank's Urbanisation Review, finalised and published in 2018, did not include any detailed city spatial-economic analysis.

12 An API integration is a connection between multiple applications that enables these systems to exchange data.

13 These delays highlight at a practical level how the much larger 'state-capture' governance events stalled operational engagements for ordinary public servants trying to get on with doing their jobs – in this case, collaborating on spatialising administrative tax data.

Mapping UIF data to PAYE tax data

In early 2017, given the delays in geocoding the PAYE tax data, the CSP and SARS tried an alternative approach. They subsequently approached the UIF about the possibility of drawing on the workplace address field in the UIF employer monthly return declaration (the UI-19 form) with the aim of analysing the spatial distribution of (formal) jobs as recorded by the UIF nationwide.

The UIF agreed to provide a full list of UIF field areas and totals to enable SARS to develop a formal specification request that would see SARS match and analyse 50 000 UIF records with PAYE records (matched by individual national identification numbers). The UIF's lengthy delays in extracting the sample data were compounded by extraction errors that excluded the requested breakdowns on age, sector, gender and remuneration. These categories were crucial for successfully matching PAYE and UIF data.

The initial SARS analytical report confirmed the potential value of integrating larger extracts of UIF data with IRP5 (PAYE) tax certificates based on individual identifiers. In principle, a picture of an individual's periods of employment and unemployment could be constructed and any inconsistencies at the individual level identified. In addition, it was hoped that the UIF data could provide the geographical workplace address of employees and thereby solve the head-office problem in the SARS data.

SARS made several recommendations to improve the UIF sample data: (1) the UIF data should contain all individuals (using monthly declaration data) and not only focus on those who had submitted UIF claims; and (2) SARS and the UIF should work with the CSP and StatsSA to standardise the UIF and PAYE tax datasets by, for instance, converting monthly to yearly remuneration in UIF data and ensuring the aggregation of income components in both datasets.

Following receipt of the revised UIF data extract, SARS explored whether the UIF monthly declaration addresses could minimise the challenge of single head-office addresses in the PAYE tax data records. However, analysis of a few large employers, particularly in the retail and banking sectors, confirmed that the majority of large employers only provide a head-office address in their monthly UIF declarations. Consequently, continued engagement with large employers at multi-establishment companies would be required to ensure greater compliance in the completion of employee main place of work address for both UIF and PAYE tax records.

The CSP and SARS agreed that the next steps in the UIF and PAYE tax data matching project would be, first, to work with the UIF team to insert an employee main place of work address field in the U-filing system (then under development and subsequently launched in March 2019); and second, to continue to work with SARS to ensure greater employer compliance in completing the employee main place of address field in the PAYE tax reconciliation submissions. However, the CSP did not pursue further engagement on the matter given that these developments were already too late for the Urbanisation Review.

Photograph by Serena Coetzee

Photograph by SDecoret / Adobe Stock

Establishing a secure administrative data centre

Establishing a secure administrative data centre

Despite the relatively inconclusive effort to obtain anonymised and geocoded tax data through engagements with SARS and the UIF for use in the Urbanisation Review, there were significant upsides to the CSP's collaboration with SARS. The work they undertook together fed a spatial element into a broader administrative data research project between SARS and the National Treasury under the Southern Africa–Towards Inclusive Economic Development (SA-TIED) programme. This programme was conducted in collaboration with UNU-WIDER over a three-year period from 2017 to 2020.

SA-TIED was motivated by the potential of tax administrative data to benefit research on economic and tax policy, wealth and income distribution dynamics, as well as on firm-level and business activity. The programme drew on the National Treasury-SARS pilot use of tax administrative data for firm-level studies using the SARS-NT Panel data, where, with the support of UNU-WIDER,[14] IT infrastructure had been provided to enable researchers to undertake previously impossible anonymised firm-level tax analysis.

Pilot project

An initial pilot project followed the international best practice of using anonymised tax and trade records for micro-level research and analysis to (1) develop a better understanding of firm dynamics and (2) identify existing constraints to growth and opportunities for economic transformation – with extremely positive results (Pieterse et al., 2016).

The pilot project enabled access to an integrated set of anonymised tax and trade records through a secure data facility at the National Treasury. The aim was to establish a secure database of anonymised tax and customs administrative data within the National Treasury and, over time, to institutionalise the database and capability within SARS itself.

The National Treasury and UNU-WIDER team first issued a series of calls for research proposals, which were evaluated according to their feasibility and policy relevance. Approved researchers were provided access to data within the National Treasury under secure conditions. There were also multiple layers of protection of the confidentiality of taxpayer data. Obvious identifiers, such as names and trading names, had been removed, and other recognisable identifiers, such as national identification numbers and tax reference numbers, had been replaced with non-intelligent identifiers. Researchers signed confidentiality agreements and were not able to remove data from the system. To eliminate the risk of indirectly identifying taxpayers, all results generated from analysis were checked before being released.

Key lessons from the pilot project indicated that: (1) dedicated administrative support is required for running a secure data centre; (2) data documentation is also required, particularly in support of lineage metadata, and field definitions and distinctions; and (3) most importantly, processing on a server is preferable to desktop processing because the datasets are so large.

Discussions held to review the pilot project drew attention to the benefits that a public economic administrative tax data centre holds in providing

14 UNU-WIDER provided the IT infrastructure for the pilot project, and dedicated staff were appointed to undertake the data manipulation and integration using anonymous data supplied by SARS.

an evidence base for tax, fiscal and economic policy impact analysis and policy formulation in line with the creation of a capable developmental state. Further benefits highlighted in the pilot review included enabling the continuous improvement of the quality of administrative data based on insights gained from a deep interrogation of tax and customs records.

Such beneficial approaches to using administrative data would achieve significant cost-savings to the state by minimising the need for expensive surveys. The pilot also suggested that, over time, a secure public administrative data centre could streamline access to the data required by academics and applied researchers in economics, and especially economic geography, as well as build analytical capacity in policy research and planning units in the public sector.

Second phase

SARS emphasised that setting up an administrative economic research data centre would involve: (1) acquiring adequate IT infrastructure for the storage, processing power and analytical tools appropriate for working with large datasets; (2) compiling an annually refreshed set of data that is cleaned, integrated and documented to provide a rich picture of the economic behaviour and circumstances of businesses and individuals; and (3) staffing the facility, ensuring that experienced researchers and people who have managed administrative data and secure data facilities are paired with staff who understand tax administration.

Given the successes of the pilot collaboration with UNU-WIDER, the National Treasury and SARS entered into a further collaboration under the SA-TIED programme. The workstreams on 'Enterprise development for job creation and growth' and 'Public revenue mobilisation for inclusive development', in particular, depended on advancing the provision of reliable tax and customs data for researchers and policy-makers while at the same time improving the quality of administrative data collection.

The workstream proposals for the second phase of the UNU-WIDER collaboration did not initially envisage a continuation of the pilot project's access to an integrated set of anonymised tax and trade records through the secure data centre at the National Treasury. Rather, it was thought that, by early 2017, SARS would have established its own secure administrative data and research facility with the necessary dedicated hardware, server storage capacity and software, as well as the dedicated staff required to develop and maintain the database and support research access to the data. In addition, the initial phase of setting up the facility would require sourcing the specialist skills and knowledge of researchers experienced in managing and using administrative data as well as their having specialist skills in econometric modelling. Securing such competencies would build SARS's internal data and research capacity for managing the facility in the medium term.

As SARS does not collect tax data for the purpose of economic research and analysis, and since SARS had other funding priorities due to a revenue crisis at that time, the National Treasury undertook to establish a more formalised economic administrative data centre within itself, and for the centre to include city-level spatial economic data that would advance the CSP's analytical support to cities. Lengthy and complex procurement processes meant that the official National Treasury Secure Data Facility (NT-SDF) only became operational in 2019.

Using administrative data would achieve significant cost-savings by minimising the need for expensive surveys

Geocoding tax data from postal codes

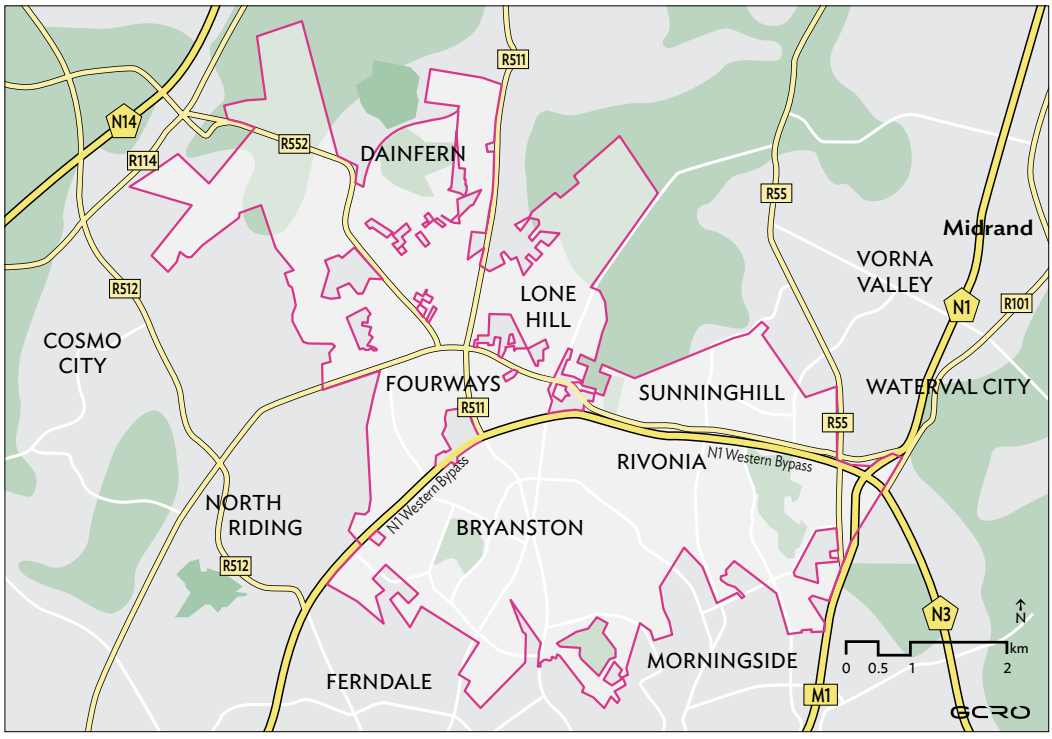
The NT-SDF strengthened collaboration with SARS under the SA-TIED programme. When the National Treasury received the 2016 tax data drawn from SARS, the latter agreed that the National Treasury would undertake the geocoding of the tax data for the SARS-NT Panel (both firm-level and individual-level panels) using the postal code in the address fields of the IRP5 tax data.

The National Treasury undertook a basic geocoding of the address data on the IRP5 tax database with external technical assistance. In brief, the postal codes for the personal and business addresses were mapped onto administrative geographic areas at four levels of geographic place types: province, district municipality, local municipality and census main-place. This is not ideal since postal codes are designed for postal delivery and not for accurate location information, rendering their usage insufficient for accurate geocoding. Furthermore, South African postal addresses often define large geographic regions, increasing the degree of geocoding inaccuracy.

A more precise geocoding of address data, one that is structured and standardised according to SANS-1883 (see fn 10), would be the preferred solution for geocoding tax data in a way that preserves the anonymity of the dataset. These arguments are clearly set out in a recent GCRO publication on spatial data

Figure 7: Postal codes represent a practical though sub-optimal geocoding solution – in this Johannesburg example, postal code 2191 covers multiple sub-places

SOURCE: Google Maps; map redrawn by Janet Alexander



infrastructure (SDI), which concludes that 'geospatial data that is maintained and provided through an SDI provides the backbone for city governance' (Coetzee et al., 2020, p. 20).

That said, the National Treasury decided that using the postal code was the best approach for geocoding the tax data given the lack of structured and standardised address data in the IRP5 tax dataset at that point. Accordingly, the National Treasury-led geocoding process linked the postal code to geographical aggregations – province, municipality, district municipality and census main-place – to facilitate spatial research while preserving the anonymity of individuals and firms. However, as postal code geography has never been used for the purposes of demographic or census data, it does not coincide with StatsSA's geographical frame and does not have associated shape files.[15]

In the absence of postal code shape files, the Google Maps (or the OpenStreetMap) API was used to calculate the geographic midpoint of each postal code and to find a geographic structure in which the postal code is located. This approach makes the following assumptions: (1) the postal codes reported by Google Maps are correct; (2) the midpoint of the postal code is an accurate reflection of the area of the postal code; and (3) the postal code does not straddle multiple geographic structures, and that the geographical structure in which the midpoint of the postal code is located is also the geographical structure that has the largest overlap with the said postal code.

There are approximately 3 400 unique postal codes recorded from the IRP5 tax data. A list of all the unique postal codes in the data was created, and each of the postal codes was passed to the Google Maps API, which returned the GPS coordinates of the midpoint of the postal codes. A file with the postal codes and their longitude and latitude was then stored.

The shape files of the various geographic structures – province, municipality and census main-place – were read in from the census data. Using the longitude and latitude coordinates, the location was mapped to the geographic structure in which the GPS coordinates fell, and the name of this structure was appended to the postal code. Due to the difficulties of using postal codes for geocoding purposes, and also to ensure anonymisation of the data, the decision was made to geocode to the census main-place as the lowest level of disaggregation, rather than to the sub-place.

This process generated a conjunction table including postal codes, provinces, district municipalities, local municipalities and census main-places. The postal codes in the IRP5 tax data were then linked to the conjunction table.

Postal code limitations

As noted above, postal codes are not ideal for geocoding because they are not designed for accurate location information and, in South Africa, often cover large geographic regions, increasing the degree of geocoding inaccuracy. National Treasury requested the SARS geocoded tax data as part of its data extraction; although, given the weaknesses in the SARS geocoding algorithm, the results may not be significantly more accurate than geocoding using the postal codes. This inconvenient fact presents further challenges to ensuring greater accuracy in the geocoding of tax data for the purposes of city-level spatial-economic research and analysis.

Even once the geocoding algorithm is improved, it would only be useful for geocoding the employee work addresses provided by small and medium-sized firms. Large multi-enterprise firms submit their employer reconciliations in batches rather than in single entries, providing the head-office address as the employee work address. This means that a real solution to the geocoding challenge requires far more extensive collaboration with SARS to work with large multi-establishment businesses to provide accurate employee work physical addresses, in addition to head-office addresses, on tax certificates such as employee IRP5s.

15 A shapefile represents a group of polygons on a coordinate space. A postal-code shapefile would then represent a map of South Africa where the unit of account is the postal code. In other words, the postal-code shapefile would define the geographical boundaries of all postal codes.

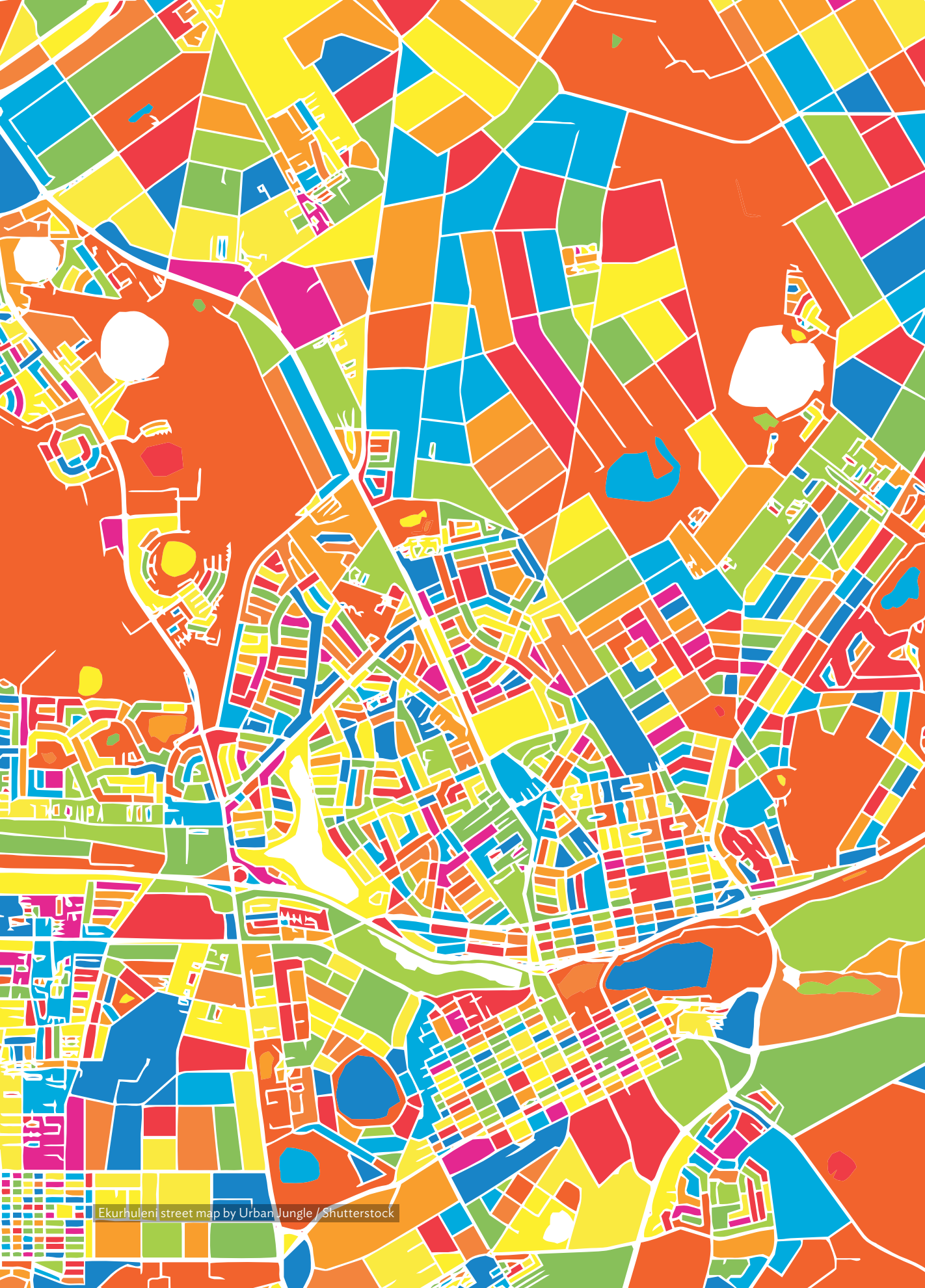
Looking ahead

Such a collaboration may now be possible given the change in leadership at SARS in recent years. Since joining SARS in mid-2019, Commissioner Kieswetter has consistently emphasised the importance of strengthening technological capacity at SARS. Recent advances in this regard include auto-population tax returns with third-party data through collaborations with financial institutions, retirement funds and medical schemes. SARS is extending these positive developments into working with 'interrogation algorithms' and 'predictive analytics' to improve the overall integrity of the national tax collection effort (Mzekandaba, 2021).

However, while SARS has a strong requirement for appropriately structured and geocoded data that is linked to their revenue collection efforts for individuals and businesses, there is not yet clear evidence that their geocoding capabilities have indeed been strengthened, or that they have the capacity for far more extensive collaboration with the National Treasury to encourage large multi-establishment businesses to provide accurate employee work physical addresses on tax certificates. This meant that the *2021 City Spatial Economic Data Reports* that the CSP published in 2021 drew on the city spatial economic resource that had been shaped to a point, but which still faces notable technical challenges related to the aggregated 'head-office' effect. It should be noted that the author of this ethnographic account was no longer working for the CSP by the time the reports were published.

These challenges notwithstanding, the successful establishment of the NT-SDF provided the level of secure IT infrastructure required to advance the SA-TIED project, and to significantly advance the agenda to use anonymised and geocoded micro tax data for broader and more focused economic research and analysis. The reports on city spatial economic data released in 2021 reflect these advances. Irrespective of the enduring limitations in the data, they present important visual analyses of employment and firm-level economic activity at the intra-city, main-place level.

Photograph by Clive Hassall


Ekurhuleni street map by Urban Jungle / Shutterstock

Conclusions, resolutions and sequels

Conclusions, resolutions and sequels

A few years further along, as we emerge from the challenges of the 'state-capture' years, South Africa's public institutions, notably SARS, are slowly being rebuilt and their institutional integrity and coherence are being restored one 'byte' at a time. As we reflect back and look forward, we should ask what insights this ethnographic account entails for understanding the assumptions underpinning the *2021 City Spatialised Economic Data Reports*. Are these insights important for taking the necessary steps to improve the integrity of city economic data and enhance their use in credible, evidence-based urban economic analysis?

Our insights draw, first, on broader institutional and public management concerns that frame the governance environment on which steps to improve city spatial economic data depend; and second, the insights draw on more data-specific conclusions which tell us that, despite the long journey travelled thus far, the original data conundrums facing the collection of city spatial economic data remain.

Finally, this paper points to adventures that will be possible in the future once the current governance and data challenges are resolved.

Improving the governance environment

Data collection depends on people and processes
This paper's narrative account stands as testament to the fact that while yielding rewards in the long term, accessing and improving the quality of intra-city spatial economic data is a long and demanding task. Primarily a people- rather than a data-driven process, it requires cultivating and maintaining good relationships with those public-sector officials who are the custodians ('owners') of the administrative processes collecting the required data.

Since the data is collected as a by-product of public administrative processes, there is a need to meet with and listen to the public officials and others whose daily work routines involve processing the data. Having an ethnographic ear helps us to appreciate the logic and detail of the public administrative processes that collect the data – which are generally aligned with the original bureaucratic purposes that order our daily lives rather than being collected for research purposes. Here, we think of SARS collecting taxes from individuals and companies, and the UIF processing employers' monthly UIF declarations and individuals' UIF claims. We need to develop strong productive interfaces with these data custodians to understand how we can re-design or add to processes *at their source* in a way that enables access to quality data for research purposes while not undermining the credibility of their original public administrative objectives.

Unlocking and improving city economic data therefore calls for champions or deal-makers who can proactively shape an administrative data learning network, building relationships at institutional, group and individual levels. These roles call for formal collaborations, as well as informal conversations at differing levels of institutional seniority, that broker and manage relationships at the same time as engaging in technical content. Keeping momentum in a learning network, and in project implementation across institutions, requires a structured and tenacious approach that creates a culture of learning and excellence at the implementation level of the system. The environment must be open, enabling transparency and a spirit of camaraderie that catalyses conversation and debate, encourages enthusiasm and opportunity in discovery, and directs diverse conversations towards positive outcomes.

Robust public institutions matter

Perhaps the most compelling overarching insight is that the quality of public institutions and their governance processes is critical to improving city economic data. And it is the people and office-bearers in institutions administering the public services underpinning our daily lives, providing the institutional foundations for South Africa's future development, who are the custodians of the administrative data we seek for research purposes. As evidenced by the governance, capacity and capability crises at SARS and the UIF in recent years, if we want good quality administrative data, the public institutions that hold this data have to be stabilised and maintained.

One cannot but bemoan the hollowing out of critical institutions in South Africa's public sector over the last decade and, for many institutions, even longer. The lack of capacity and, in many cases, capability itself, compounds the transgression of public interest by individual and vested interests. Our narrative shows that administrative data champions have to be tenacious and resourceful: practising patience for the long game rather than focusing on short-term objectives; and devising innovative solutions where initial opportunities or pathways close. In this respect, it is important, first, that the heads of public institutions as well as their technical teams are part of the data conversation to ensure that such dialogue is as collaborative and openly dynamic as possible; and second, that these personnel show leadership by taking responsibility for public institutional performance and accountability.

A courageous further step would be to set a specific agenda – led by the National Treasury, SARS and StatsSA, together with the South African Information Regulator – for using linked anonymised public-sector administrative economic data within the ambit of the Protection of Personal Information Act (No. 4 of 2013). Such a focused and strategic approach is key to addressing the underlying challenges that still confound the collection of city economic data.

Fixing data collection conundrums

The 'head-office effect' is still here

A principal reason for pursuing access to geocoded and anonymised tax administrative data was to overcome the main challenge of the 'head-office effect'. The *2021 City Spatialised Economic Data Reports* introduce the key innovation of mapping geocoded and anonymised tax data by Uber's H3 spatial index. This consists of equal area hexagons with various sizes into which data – previously geocoded by large, unwieldy postal code areas – can more accurately be distributed. While this is a significant enhancement, it does not address the underlying 'head-office effect'. The data still retains a head-office bias given that it is based mainly on IRP5 tax certificates, where many large multi-establishment companies still submit batch returns that state their head office in the address field, without adding the main place of work for each individual employee.

The problem is acknowledged in the release notes accompanying the applications forms for accessing the data. These notes suggest that the challenge may be fairly limited, but note that the exact scale of the problem remains unknown, with more research required to determine the actual extent of any potential head-office bias (Nell and Visagie, 2022).

Despite the significant geocoding enhancements made, the 'head-office effect' in the data can only really be countered at source. Next steps will require further collaboration with SARS to accurately weigh the scale of the challenge and correct for it at the point at which data for large multi-establishment firms is entered.

Industry classification is imputed

The *2021 City Spatialised Economic Data Reports* present the economic or industry classification of the data at SIC 7 1–5-digit level, but do not mention that these industry classification data points are largely imputed.

The quality of public institutions and their governance processes is critical to improving city economic data

Budlender and Ebrahim (2020) cover the complexity of the four usable sources of industry classification in the SARS-NT data panel in detail. They conclude that the sources 'vary in their firm coverage, comprise different industry classification systems, and often assign conflicting industries to any given firm' (p. 1). A detailed read of Budlender and Ebrahim (2020) is enlightening and facilitates an understanding of the challenges in imputing industry classification to the tax data sources.

The current lack of a consistent system of industry/economic classification for underlying tax data draws into question the presentation of the city spatial economic data at SIC 5 5-digit level. It is likely that the imputations at this level of disaggregation lose significant credibility, and that higher levels of aggregation to 2-digit level, for instance, may present more credible results.

The National Treasury has not yet released the detailed metadata document identifying each variable and the information related to these variables in city economic data; nor have they released the detailed methodological document highlighting the process by which the data was prepared and the limitations associated with this process. These detailed metadata and methodological documents are critical for understanding the limitations that underpin the imputation of economic/industrial classification to the underlying tax data source. Researchers need to refer to these metadata and documents in their utilisation of the processed city economic data for their research.

Future adventures

Professionalising data science

This GCRO Occasional Paper highlights the incredible potential that administrative economic data holds for creating a robust evidence base. This potential applies not only to South Africa but also more broadly to developing economies worldwide that do not have the means to produce an official annual economic survey or census at establishment level, and where a robust evidence base is critical for responding to rapid urbanisation trends.

Data science is an emerging vocation, and data-driven approaches are increasingly reshaping the way urbanists analyse and make policy decisions in city space. Using data science to understand urban opportunities and challenges is gaining traction worldwide, leading to the development of a new field – urban science (Duarte and deSouza, 2021). Cities and the public sector are being urged to invest in developing these scarce skills given the significant potential urban science holds for reforming urban spaces, improving their ease of use and providing opportunities for people to live, work and play. Urban data and the emerging praxis of urban science are

Photograph by André Branco

becoming key levers for cities to develop a robust evidence base that underpins planning, budgeting and delivering quality services to an increasingly well-informed and digital-savvy urban population.

Casting data champions

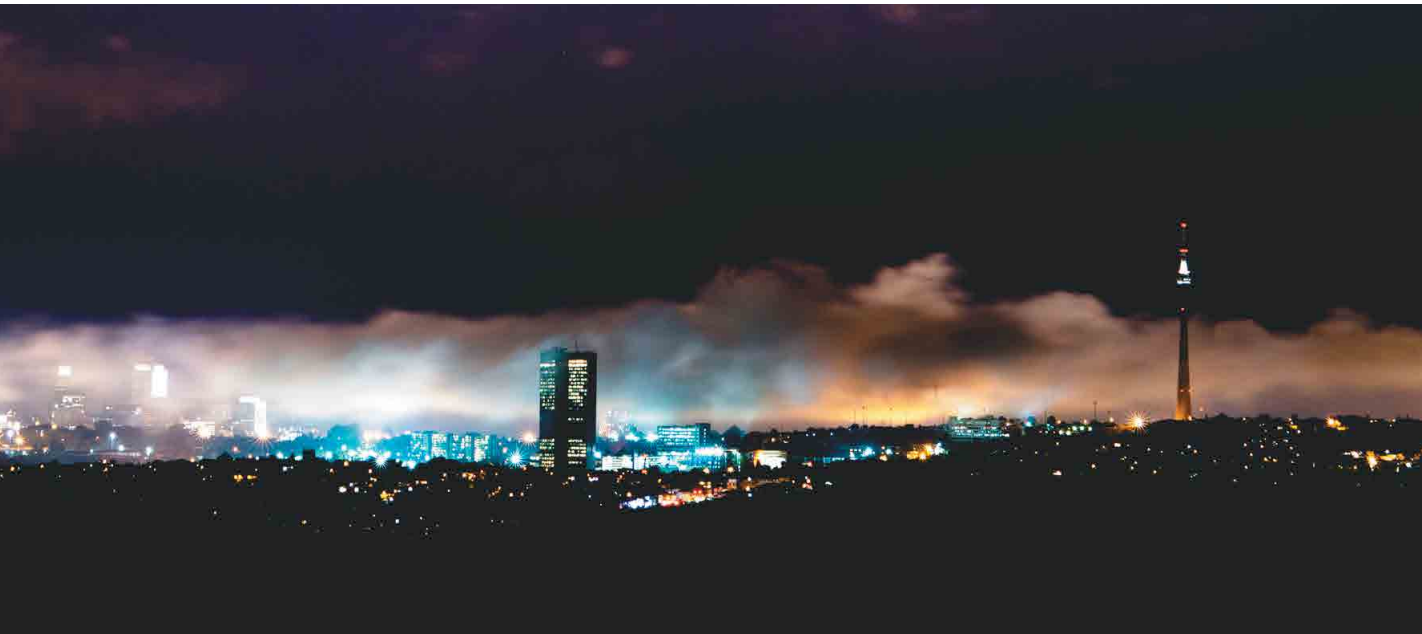
This paper throws into sharp relief the immense amount of dedicated person-energy that is required to build a robust administrative data learning network from which data for research purposes may be unlocked. Given the nature of administrative datasets themselves, this energy is primarily directed at engaging with people-led rather than data-led processes.

We need a number of champions, working together in a coordinated institutional effort, to realise the potential of administrative data to underpin an urban evidence base. Scaling up this effort requires the discipline of data science to mature into professional career paths within the public sector. In doing so, urban science holds enormous potential for underpinning significant improvements in city economic performance by providing a more robust evidence base for urban research as well as for policy formulation and decision-making.

The timing is right for this move. The work of this author took place in a time and space severely constrained by 'state-capture' rent-seeking, which weakened institutional coherence and capability. Creative space is now opening. That said, charting innovative ways in government requires that new approaches be tested in an open space that is not defined or constrained by bureaucratic standardisation and strictures. Creative spaces require a level of seniority that is able to hold and promote a vision and to test new approaches for success and upscaling in implementation. The President's Economic Advisory Council and the Presidency's Project Management Office offer such a dynamic environment. Underpinned by a robust urban-science evidence base, a renewed urban agenda has the potential to make a significant contribution to the government's Economic Reconstruction and Recovery Plan, similar to the innovative lens framing the Presidential Economic Stimulus public employment approaches in response to the COVID-19 pandemic.

Reaching beyond South Africa

Finally, there is an opportunity for this journey to extend beyond South Africa's borders as capacity in urban and tax administration strengthens in other developing countries due to continued advances in technology and automation. Think of an African urban administrative data network. Leap further to imagine a Global South urban administrative data network. The possibilities of such developments are endless as well as exciting in terms of their potential rewards.



References

ATAF (African Tax Administrative Forum). (2021). *Annual Report 2020*.

Bahl, R., Smoke, P. and Solomon, D. (2003). Overview of the local government revenue system. In R. Bahl and P. Smoke (Eds.), *Restructuring local government finance in developing countries: Lessons from South Africa* (pp. 71–91). Cheltenham: Edward Elgar Publishing.

Boyd, C. (2022). Data as assemblage. *Journal of Documentation*. https://doi.org/10.1108/jd-08-2021-0159

Budlender, J. and Ebrahim, A. (2020). *Industry classification in the South African tax microdata*. SA-TIED Working Paper No. 134. Southern Africa–Towards Inclusive Economic Development. https://sa-tied.wider.unu.edu/sites/default/files/pdf/SA-TIED-WP-134.pdf

CAFH (Centre for Affordable Housing Finance in Africa). (2021). *eThekwini housing market report – 2021*. https://housingfinanceafrica.org/app/uploads/2022/02/eThekwini-Property-Report-2021.pdf

City of Cape Town. (2021). *Economic Performance Indicators for Cape Town (EPIC): 2021 Quarter 1*. https://resource.capetown.gov.za/documentcentre/Documents/City%20research%20reports%20and%20review/CCT_EPIC_2021_Q1.pdf

CoGTA (Department of Cooperative Governance and Traditional Affairs, Republic of South Africa). (2016). *Integrated Urban Development Framework (IUDF)*. https://www.africancentreforcities.net/wp-content/uploads/2017/05/IUDF-2016_WEB-min.pdf

Crankshaw, O. (2022). *Urban inequality: Theory evidence and method in Johannesburg*. London: Zed.

CSIR (Council for Scientific and Industrial Research). (2020). *SA CSIR MesoZone 2018v2 Dataset*. http://stepsa.org/socio_econ.html#Geospatial

CSP (Cities Support Programme). (2021). *City Spatialised Economic Data Reports*. Pretoria: National Treasury. http://www.governmentpublications.lib.uct.ac.za/news/2021-city-spatialised-economic-data-reports

Duarte, F. and deSouza, P. (2021). Data science and cities: A critical approach. *Harvard Data Science Review*, *2*(3). https://doi.org/10.1162/99608f92.b3fc5cc8

ERLN (Economies of Regions Learning Network). (2015). Meeting agenda, 19–20 August. https://erln.gtac.gov.za/images/jevents/55fffbd4edd4f8.98203578.pdf

ERLN (Economies of Regions Learning Network). (2016). Data Technical Working Group meeting, 27–28 July. https://erln.gtac.gov.za/erln-learning-events/past-events/eventdetail/42/-/data-technical-working-group-meeting

eThekwini Municipality. (2021). *eThekwini economy 'at a glance': February 2021*. https://edge.durban/dataset/b51325fd-ef89-42b8-8b83-b8f3f324d3b7/resource/74fdbdfe-2a10-4977-b04f-fb35eb83932c/download/economy-at-a-glance-february-2021.pdf

FFC (Financial and Fiscal Commission). (2013). *Sustaining local government finances: Final report on the Financial and Fiscal Commission's public hearings on the review of the Local Government Fiscal Framework*. https://www.gov.za/sites/default/files/gcis_document/201409/financialandfiscalcommissionsustaining localgovernmentfinancesreport.pdf

Gavin, E., Breytenbach, D., Carolissen, R. and Leola, M. (2013). The role of tax administrative data in the production of official statistics in South Africa (STS021). In *Proceedings 59th ISI World Statistics Congress* (pp. 1483–1488), Hong Kong, 25–30 August 2013. https://2013.isiproceedings.org/Files/STS021-P2-S.pdf

Goswami, A.G. and Lall, S.V. (2016). *Jobs in the city: Explaining urban spatial structure in Kampala*. Policy Research Working Paper No. 7655. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/24230

Götz, G. and Todes, A. (2014). Johannesburg's urban space economy. In P. Harrison, G. Götz, A. Todes and C. Wray (Eds.), *Changing space, changing city: Johannesburg after apartheid* (pp. 117–136). Johannesburg: Wits University Press.

Harrison, P. (2020). *Johannesburg and its epidemics: Can we learn from history?* GCRO Occasional Paper No. 16. Johannesburg: Gauteng City-Region Observatory. https://cdn.gcro.ac.za/media/documents/GCRO_Occasional_Paper-Epidemics.pdf

Harrison, P., Götz, G., Todes, A. and Wray, C. (2014). Materialities, subjectivities and spatial transformation in Johannesburg. In P. Harrison, G. Götz, A. Todes and C. Wray (Eds.), *Changing space, changing city: Johannesburg after apartheid* (pp. 2–39). Johannesburg: Wits University Press.

Isaacs, E. (2013, 1 March). *Ethnography*. Ellen Isaacs at TEDxBroadway. https://www.youtube.com/watch?v=nV0jY5VgymI

Kaziboni, L., Mondliwa, P. and Robb, G. (2015). Towards an understanding of the economy of Johannesburg: Industrial Nodes Report. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.2716031

Lall, S.V., Henderson, J.V. and Venables, A.J. (2017). *Africa's cities. Opening doors to the world*. Washington, DC: World Bank. https://documents1.worldbank.org/curated/en/854221490781543956/pdf/113851-PUB-PUBLIC-PUBDATE-2-9-2017.pdf

Mabin, A. (2013). *The map of Gauteng: Evolution of a city-region in concept and plan*. GCRO Occasional Paper No. 5. Johannesburg: Gauteng City-Region Observatory. https://cdn.gcro.ac.za/media/documents/gcro_occasional_paper_5_-_mabin_map_of_gauteng_july_2013_final.pdf

Malmidin, J. (2017). *Output from administrative data as the basis for decisions*. Presentation at the United Nations World Data Forum, Cape Town, 15–16 January.

McCarthy, T. (2010). *The decanting of acid mine water in the Gauteng City-Region*. GCRO Provocation. Johannesburg: Gauteng City-Region Observatory. https://cdn.gcro.ac.za/media/documents/gcro_terence_mccarthy_amd__final_version.pdf

Mzekandaba, S. (2021, 5 July). SARS pins hopes on key tech as modernisation plans advance. *ITWeb*. https://www.itweb.co.za/content/VgZey7JoZGjMdjX9

Naidoo, Y. (2019). *Industrial and commercial buildings*. GCRO Map of the Month, July 2019. Johannesburg: Gauteng City-Region Observatory. https://cdn.gcro.ac.za/media/documents/MoTM_2017.07_Industrial_and_commercial.pdf

National Treasury. (2009). *Response to unfounded claims by the City of Cape Town of unfair allocations to the City by the National Treasury*. Pretoria: National Treasury. http://www.treasury.gov.za/comm_media/press/2009/2009022001.pdf

National Treasury. (2018a). *Guidance note for the formulation of Built Environment Performance Plans (BEPPs)*. Pretoria: National Treasury. https://csp.treasury.gov.za/Projectdocuments/BEPP%20Guidelines%202016_17.pdf

National Treasury. (2018b). *Managing urbanisation to achieve inclusive growth. A review of trends in South African urbanisation and suggestions for improved management of urbanisation.* Pretoria: National Treasury. https://csp.treasury.gov.za/csp/DocumentsProjects/Managing%20Urbanisation.pdf

Nell, A. and Visagie, J. (2022). Spatial Tax Panel, 2014–2021. Release notes: Version 2 (14 October 2022). Pretoria: Human Sciences Research Council and National Treasury.

Nugent, R. (2018). *Commission of inquiry into tax administration and governance by SARS report*. Pretoria: The Presidency. https://www.thepresidency.gov.za/sites/default/files/SARS%20Commission%20Final%20Report.pdf

Pieterse, D., Kreuser, C.F. and Gavin, E. (2016). *Introduction to the South African Revenue Service and National Treasury firm-level panel*. UNU-WIDER Working Paper No. 2016/42. Helsinki: United Nations University World Institute for Development Economics Research. https://www.wider.unu.edu/sites/default/files/wp2016-42.pdf

Republic of South Africa (Department of Rural Development and Land Reform and Department of Planning, Monitoring and Evaluation). (2020). *Draft National Spatial Development Framework*. https://www.gov.za/sites/default/files/gcis_document/202002/draftnsdf20jan2020-compressed.pdf

Ruotsalainen, K. (2017, 17 January). *Use of administrative data: Nordic experiences*. Presentation at the United Nations World Data Forum, Cape Town, 15–18 January.

SACN (South African Cities Network). (2011). *State of South African cities report 2011: Towards resilient cities*. Johannesburg: SACN. https://www.sacities.net/wp-content/uploads/2020/03/SACN-2011-Report.pdf

SACN (South African Cities Network). (2016). *State of South African cities report 2016*. Johannesburg: SACN. http://www.socr.co.za/wp-content/uploads/2016/06/SoCR16-Main-Report-online.pdf

SARS (South African Revenue Service). (2021). *Business requirements specification: PAYE employer reconciliation*. Pretoria: SARS. https://www.sars.gov.za/wp-content/uploads/Docs/PAYE/BRS/SARS_PAYE_BRS-PAYE-Employer-Reconciliation_V20-0-2.pdf.

Shilpi, F., Xu, L., Behal, R. and Blankespoor, B. (2018). People on the move: Spatial mismatch and migration in post-apartheid South Africa. *World Bank Urbanisation Review*, Paper No. 1. https://csp.treasury.gov.za/csp/DocumentsToolbox/Paper%201%20-%20People%20on%20the%20Move.pdf

Sinclair-Smith, K. and Turok, I. (2012). The changing spatial economy of cities: An exploratory analysis of Cape Town. *Development Southern Africa, 29*(3), 391–417. http://dx.doi.org/10.1080/0376835X.2012.706037

StatsSA (Statistics South Africa). (2003). *Sampling methodology for economic statistics*. Pretoria: StatsSA. http://www.statssa.gov.za/publications/DiscussSamplingMeth/DiscussSamplingMeth.pdf

StatsSA (Statistics South Africa). (2011). *Census 2011 metadata*. Pretoria: StatsSA. http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Metadata.pdf

StatsSA (Statistics South Africa). (2012). *Standard Industrial Classification of All Economic Activities* (7th edition). Report No. 09-90-02. Pretoria: StatsSA. http://www.statssa.gov.za/classifications/codelists/Web_SIC7a/SIC_7_Final_Manual_Errata.pdf

StatsSA (Statistics South Africa). (2017). *Survey of the employed and self-employed 2017.* Statistical Release PO276. Pretoria: StatsSA. http://www.statssa.gov.za/publications/P0276/P02762017.pdf

Storper, M. (2013). *Keys to the city: How economics, institutions, social interaction, and politics shape development*. Princeton, NJ: Princeton University Press.

Todes, A. and Turok, I. (2018). Spatial inequalities and policies in South Africa: Place-based or people-centred? *Progress in Planning, 123*, 1–31. https://doi.org/10.1016/j.progress.2017.03.001

Turok, I. and Borel-Saladin, J. (2013). *The spatial economy: Background research report for the Integrated Urban Development Framework.* Pretoria: Department of Cooperative Governance and Traditional Affairs. https://www.cogta.gov.za/cgta_2016/wp-content/uploads/2016/05/TUROK-SPATIAL-ECONOMY-DRAFT-FINAL.pdf

Turok, I. and Borel-Saladin, J. (2014). Is urbanisation in South Africa on a sustainable trajectory? *Development Southern Africa*, *31*(5), 675–691. https://doi.org/10.1080/0376835X.2014.937524

UN DESA (United Nations Department of Economic and Social Affairs). (2008). *International Standard Industrial Classification of All Economic Activities (ISIC).* Statistical Papers Series M No./Rev.4. New York: UN Statistics Division. https://unstats.un.org/unsd/publication/seriesM/seriesm_4rev4e.pdf

UN DESA (United Nations Department of Economic and Social Affairs). (2018). *World urbanisation prospects: The 2018 revision*. New York: UN Population Division. https://population.un.org/wup/publications/Files/WUP2018-Report.pdf

WEF (World Economic Forum). (2018). *Future of jobs: Employment, skills and workforce strategy for the Fourth Industrial Revolution*. Cologny, Switzerland: WEF. http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

World Bank. (2016a). South Africa Urbanisation Review Inception Mission. Unpublished Aide Memoire.

World Bank. (2016b). *South Africa: Need for Economic Data* [PowerPoint presentation]. Background input to CSP Urbanisation Review meeting on 28 July 2016.

World Bank. (2021). *South Africa 2020 country profile*. World Bank Enterprise Surveys. Washngton, DC: World Bank. https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/country/South-Africa-2020.pdf

GAUTENG CITY-REGION OBSERVATORY

6th Floor University Corner
11 Jorissen St (Cnr Jorissen and Jan Smuts)
Braamfontein
Johannesburg
Gauteng, South Africa

tel +27 11 717 7280
email info@gcro.ac.za
www.gcro.ac.za