



SEAD SA
SPATIAL ECONOMIC
ACTIVITY DATA
South Africa

WORKSHOP SUMMARY

‘What data is needed for cities to thrive?’

Insights from the United Kingdom’s (UK’s)
Office for National Statistics

Held at Isbalo House, Statistics South Africa, Pretoria
on 12-13 March 2024

Version 1: June 2024

Author: Mark Paterson, senior journalist and communications consultant to the HSRC

OVERVIEW

A two-day workshop was convened by Spatial Economic Activity Data – South Africa (SEAD-SA) with the support of the United Kingdom’s (UK’s) Foreign, Commonwealth and Development Office (FCDO) on 12-13 March 2024 to identify and promote the provision of the kinds of data required by municipal authorities to foster inclusive, vibrant economies in South African cities.

The meeting brought together over 150 civil servants, researchers and practitioners to establish and widen a “Community of Practice” producing and using administrative data so that economic decision-making may increasingly be evidence-based. To this end, there was a focus on learning from the efforts of the UK’s Office for National Statistics (ONS) to provide and improve the standard of spatial economic data available to local authorities in the UK.

The workshop was convened by the Human Sciences Research Council (HSRC); National Treasury; and the UK-FCDO in partnership with the metros; the South African Revenue Service (SARS); the United Nations University World Institute for Development Economics Research (UNU-WIDER); the South African Local Government Association (SALGA); the South African Cities Network (SACN); and the University of the Free State (UFS).

The meeting sought to:

- Identify gaps in the South African system for producing data in support of economic decision-making;
- Offer inspiration and provide training, learning from the efforts of the ONS;
- Share the UK’s approach to evaluating the impacts of policy on local economies;
- Provide case studies from eThekweni and Cape Town on producing economic data for decision-making;
- Introduce the Spatial Tax Panel produced by SEAD-SA; and
- Brainstorm future topics and priorities for the SEAD-SA Community of Practice.

Under the tagline, “Data for inclusive and vibrant city economies”, the SEAD-SA Community of Practice seeks to generate new knowledge and an exchange of ideas alongside better practice in strengthening the performance of South African cities. The purpose is to broaden the circle of researchers, policymakers and practitioners interested in the economy of cities by building an active network of enthusiasts and experts.

The Community of Practice initiative takes advantage of a Secure Data Facility (SDF) established at the National Treasury with the support of SARS and managed by UNU-WIDER. Data produced at this facility has been deployed by the SEAD-SA project to establish an open-access interactive web portal built by the HSRC and hosted by UFS that aims to spread new datasets about local economic activity among municipal policy- and decision-makers (see www.spatialtaxdata.org.za).

The data currently being produced by SEAD-SA, which based on tax data provided by SARS, has already underpinned a major report on the economic outlook for South Africa’s cities produced by HSRC and Treasury (see <https://spatialtaxdata.org.za/resources>) and is being deployed and championed by local municipalities, notably the City of eThekweni and the City of Cape Town.

1. OPENING THE WORKSHOP

1.1 *The story behind SEAD-SA*¹

This present workshop marks a milestone in continued efforts to provide municipalities, researchers and other interested parties with free spatial economic data with the goal of improving public and private-sector planning, investment and monitoring.

The initiative to provide this information began with the metros championing the need for spatialised economic activity data to help them plan, budget and manage more effectively. Responding to this drive and collaborating with SARS and UNUWIDER through the “South Africa – Towards Inclusive Economic Growth” programme, the National Treasury established a Secure Data Facility in 2015. The facility housed anonymised tax data, primarily mined to influence macro-economic policy.

The metros lobbied for access to the data through an Economy of Regions Learning Network (ERLN) convened by the Government Technical Advisory Centre (GTAC) and subsequently through National Treasury’s Cities Support Programme. In 2021, a consultative process to develop a spatialised economic activity data strategy for the country was initiated with funding from the Swiss State Secretariat for Economic Affairs (SECO) – and spatialising the anonymised tax data in the Secure Data Facility was identified as a quick win.

Andrew Nell, a data scientist, was contracted to mine the tax data and, in May 2021, eight metro spatialised economic activity reports were launched as a result of collaboration between the facility and SARS. The reports were enthusiastically received by the metros, and SARS was asked if it could provide tax panel datasets so that the metros could apply and manipulate the data as they saw fit.

In July 2021, National Treasury and HSRC started collaborating to secure sustainable access to the anonymised and spatialised tax data for the metros and to enhance the capacity for the application and interpretation of this data. Under the leadership of SALGA, lobbying for access to this data mounted and, in November 2021, SARS agreed to release anonymised tax data to the municipalities in the form of prepared and de-identified panels.

In June 2023, SEAD-SA, which is jointly led by HSRC, the metros, National Treasury, SALGA and the SACN, was launched with the support of the national ministers of finance, trade and investment, and co-operative governance. SEAD-SA offers a Spatial Tax Data Portal, managed by the HSRC and UFS, and produces an annual City Economic Outlook.

Metros led by eThekweni and Cape Town have used and applied the data in innovative ways and have demonstrated their willingness to share their approaches and systems with others. In addition, the data has now been integrated into local government reporting reforms through Circular 88, which allows government to monitor the outcome of their investments and programmes on formal employment levels and numbers of formal micro and small businesses.

IN 2023, the UK-FCDO agreed to provide financial and technical support for SEAD-SA, which led to the present partnership with the UK’s Office of National Statistics.

SEAD-SA is a good new story. The initiative has catalysed change and has demonstrated the potential that lies in the social, economic and financial datasets that are held across government if they can be spatialised and made available at a local level. In this respect, National Treasury is committed to building on the experience of SEAD-SA and collaborating to realise the larger vision of a South African “integrated data lake.”

¹ This address was delivered by Malijeng Ngqaleni, Deputy Director-General: Intergovernmental Relations, National Treasury.

SEAD-SA has shown that change takes time – time to build trust and confidence; time to pilot and demonstrate success; and time to include everyone – but that, if time is given, the results can be impactful.

1.2 Sound, accountable data use can create great benefits²

Data becomes increasingly valuable, the more granular and the more local it is. In this respect, the current workshop presents an opportunity to learn and explore how massive amounts of micro data can be used to make a huge, transformational difference to the lives of citizens. At the same time, it is important to be careful about how such data is used. It is important to ensure that the data being deployed is sound because otherwise the conclusions that are produced will be nonsensical. In addition, the ways in which the data is analysed must be sound in order to ensure that the conclusions are thoughtful and the evidence-based policymaking that is produced as a result is effective. It is also important to ensure that there is transparency and accountability throughout the process in terms of how the anonymised data is accessed and manipulated. If citizens fail to trust those producing and using the data and the ways in which the data is being used, then they may undermine the provision of the reliable data on which effective government administration depends. In the UK, the work undertaken by ONS is underpinned by a strong informational infrastructure founded on trust and integrity, as well as the knowledge and skills required to manipulate and draw conclusions from the data properly. These are also understood to be the foundational elements of the data ecosystem in South Africa – and it is this shared understanding which underpins the present partnership in this area between the two countries.

1.3 Risks and benefits of using administrative data³

Stats SA produces monthly import and export unit-value indices, and information on liquidations and insolvencies based solely on administrative data; and also uses administrative data as the foundation for its sample surveys and to provide sample frames for monthly data sets on manufacturing, electricity usage and the retail trade.

However, administrative and other data sources are increasingly under threat from hacking. Hackers have accessed information held by the government employee pension fund, as well as the Companies and Intellectual Property Commission (CIPC), leading to the release of a lot of data on companies and directors in the public domain. In addition, the Department of Justice, which is a primary data source for one of Stats SA's monthly series, has been hacked, leading to that series being discontinued. In this respect, a problem with relying on administrative data is that it can lead to incomplete statistics being released when the underlying data is unavailable or has been compromised. For example, consumer price index (CPI) statistics, some aspects of which rely on administrative data, may be considered unreliable when constituent datasets, for example, concerning food prices, cannot be provided. So, there is a problem with designing statistical releases that may depend on administrative data beyond the control of statisticians, notwithstanding the benefits of such data. The risks involved can make official statisticians quite nervous.

In general, administrative data must meet certain minimum standards before being used in official statistics. However, it can be difficult to ascertain whether data from external sources meets these standards. So, statistics produced using such data, however useful they may be, should perhaps be marked as “experimental”, so that their possible inaccuracy does not undermine the status of the officially sanctioned statistics being produced. In this respect, there are questions around the kinds of administrative data that may be used and around Stats SA's role in supporting the production of statistics using administrative data that are not under its control. Where should the line be drawn in terms of the credibility of administrative data? Should Stats SA be adopting a proactive role or an advisory one in relation to the use of such data?

² This address was delivered by Mike Foster, Economic Counsellor, British High Commission, Pretoria.

³ This address was delivered by Joe De Beer, Deputy Director-General: Economic Statistics, Statistics South Africa (Stats SA).

In seeking an answer to these questions, the benefits of using administrative data should be acknowledged, including in relation to:

- Cost efficiency. The use of administrative data, such as registrations of births and deaths, allows statisticians to reuse existing data instead of undertaking new surveys, which significantly reduces costs.
- Timeliness: Unlike traditional surveys which take time to implement and analyse, administrative data offers the possibility of scheduled, timely information that may be used to update data sets.
- A reduced response burden. The use of existing administrative records reduces the need for additional surveys, minimising the burden on respondent individuals and organisations.

At the same time, the use of administrative data poses significant challenges in relation to:

- Insufficient quality and consistency: The administrative data sources may vary in terms of quality due to different local practices and circumstances, inhibiting consistency and standardisation across the various data sources. The extent of the efforts required to address such inconsistency may undermine the cost-effectiveness of this mode of data production; and
- Operational limitations: The systems used to collect administrative data might not offer alignment with the statistical definitions used to organise data. Differences in data-collection procedures, the information technologies being deployed and local policies governing the kinds of information being garnered can impact negatively on the usefulness of the resulting statistics.

2. VISION FOR THE SEAD-SA COMMUNITY OF PRACTICE⁴

Insufficient attention has been paid to the economic performance of South African cities, regions and local economies, partly because of a lack credible data. Local officials are required to plan for more productive and inclusive cities, but lack information about the “what” and “where” of employment and investment in these places. In response, SEAD-SA has sought to plug this gap by leveraging tax and other administrative sources of information to produce granular spatial economic data.

Responding to the lack of local economic activity data, metros identified a number of relevant administrative data sources; and SARS, acknowledging the value of tax data as offering more than a means of promoting regulatory compliance, supported the provision of anonymised, spatialised versions of the data that it held for the Secure Data Facility housed at National Treasury. The success of this initiative led to an MoU between National Treasury and the HSRC and increased stakeholder collaboration, which, in turn, led to the establishment of a spatial tax data portal; a collaborative SEAD-SA brand; and a knowledge-sharing partnership with the UK’s ONS.

The provision of high-quality data is a necessary, but insufficient, factor in the drive to improve the economic performance of cities and regions. So, the intent of the Community of Practice is to build a vibrant network of enthusiasts and experts who can innovate, disrupt and share best practices, ultimately generating new knowledge about local economies.

It is clear that hard empirical evidence is important to show the significant contribution that cities make to the South African economy, and to foster an understanding of what makes them work effectively. For example, the concentration of employment in cities – nearly two thirds of all formal jobs in the country are located in the six biggest metros – is due in large part to the great advantages that businesses derive from proximity to their customers, suppliers and workforce. There are also enormous economies of scale in providing public infrastructure in big cities. In addition, the kind of dense human interaction that only occurs in cities promotes learning, creativity and innovation which are important for productivity and economic prosperity.

Municipalities are already taking advantage of the kind of spatial economic activity data provided by SEAD-SA, and making use of it to improve their processes. In eThekweni, location-level data is used to foster an understanding of what is taking place economically in the various wards, and why some wards are more competitive on certain metrics than others. Such data further enables the municipality to make appropriate interventions in specific locations, townships and places, including industrial areas; central business districts; semi-rural areas on the outskirts of the city; and areas that are governed by tribal authorities. Spatial economic data enables comparison among areas and facilitates the production of targeted, specific economic policies which not only foster growth but also reduce economic disparities among different parts of the city. Within the administration, such data enables location-specific, evidence-based decision-making rather than decision-making based on some loose unverifiable knowledge.

For its part, National Treasury is committed to building a stronger evidence base for economic policy- and decision-making and research, with the aim of supporting local economies and driving better value for money in policy choices, through initiatives such as the SA-TIED programme, the SDF and SEAD-SA.

Building on the work undertaken so far by the various stakeholders, the aim now is to establish a Community of Practice that can advance it further. There are three aspects to a community of practice, that is, a group of people who share a common interest or passion and come together to learn from one another, solve problems and exchange knowledge and experiences. The three aspects are: the domain – what it is that the community cares

⁴ This section is based on a presentation made by Prof Justin Visagie, Senior Research Specialist, Human Sciences Research Council (HSRC); and Associate Professor, University of the Free State (UFS).

about; the community itself – the people doing the caring; and the practice, what the community can do together about the topic at hand.

The domain for the present Community of Practice is the data that comprises the evidence base, but also the cities where most of the country's economic development takes place due in large part to the economic virtues of proximity. South Africa's cities need to thrive for the national economy to thrive. The domain also concerns the kind of economic growth being sought which should be about more than growth for growth's sake and should entail the production of equitable outcomes. All these aspects of the domain have been combined in the tagline for the Community of Practice: "Data for inclusive and vibrant city economies."

The "community" aspect of the new initiative may be illuminated by a quote from the World Bank: "Communities of practice (CoPs) exist because we realise that we live in a complex world full of adaptive challenges which no one person can address and solve on their own, in isolation. Fundamentally, CoPs are created when we realise that we don't have all the answers and that we need help – we need to reach out to others for knowledge, expertise, and experience."⁵ The accuracy of this assertion would seem to be borne out by the history of the SEAD-SA programme, which has indicated that success has been produced through collaboration among multiple stakeholders.

Turning to the issue of "practice", there are a range of activities that may be undertaken as part of the new initiative:

- Knowledge-sharing: The community could host regular webinars and/or seminars, and create an online repository for sharing ideas;
- Commissioning collaborative research: This would entail establishing stronger links with academic partners and institutions;
- Fostering networking and partnerships;
- Disseminating best practices: For example, case studies and success stories from different cities outlining different approaches to municipal problems, such as public transport, infrastructure, and investment decisions, may be shared;
- Professional development and capacity building: This could entail training sessions on data-analysis techniques, visualisation tools and data-management best practices; and
- Monitoring and evaluation (M&E): Key metrics may be developed for assessing the effectiveness of data-driven interventions in driving inclusive growth.

In support of the new initiative, a SEAD-SA Community of Practice Survey is being undertaken in an effort to ensure the responsiveness of the programme and assess views on the value of the programme, as well as key areas of interest among those who are participating or planning to participate in the community. An initial instant-response survey conducted at the present meeting identified a number of benefits that may be derived from the initiative, including: learning about a range of data initiatives and sources in support of evidence-based policy- and decision-making; the provision of data required to monitor and evaluate differently; access to more granular local data; the opportunity to learn from the best; knowledge exchange and knowledge sharing; the promotion of intra-governmental collaboration on data and information; and an opportunity to build networks and connect with like-minded individuals.

A crucial aspect of any community of practice is the level of engagement among its members which determines how effective it may be. As Wenger et al. (2002) noted in *Cultivating Communities of Practice: A Guide to*

⁵ See World Bank, Collaboration for Development, Communities4Dev, web page. Accessed at: https://collaboration.worldbank.org/content/sites/collaboration-for-development/en/groups/communities4Dev/blogs.entry.html/2021/03/24/definition_of_communityofpractice-zmku.html.

Managing Knowledge: “Because communities of practice are voluntary, what makes them successful over time is their ability to generate enough excitement, relevance and value to attract and engage members ... nothing can substitute for this sense of aliveness.” In this spirit, the aim of the current initiative is to start to improve economic outcomes in South African cities and nationally by tapping into new data sources; by building capacity; and by sharing ideas – which is a goal that should generate great enthusiasm.

3. MAKING LOCAL DATA ACCESSIBLE: THE UK'S APPROACH⁶

In the UK, it was found that local data was not granular or timely enough and was hard to use. In addition, it was quite difficult to access. For example, information on labour market statistics and gross domestic product or productivity were held on separate national databases, requiring separate searches to extrapolate locally relevant data, which anyway was not designed to be that relevant as evidence for local policy- and decision-making. However, under a recent “levelling-up” policy introduced by the national government with the aim of reducing regional economic disparities and promoting devolved governance, there has been a drive to collaborate with regional stakeholders and to identify their data needs in pursuit of more effective evidence-based economic decision-making.

In seeking to address and meet local data needs, a number of challenges have been encountered. For example, in terms of the quality and kind of data being made available, there can be issues of disclosure and commercial sensitivity concerning the economic advantages that may be derived by firms occupying a dominant position in the market. Meanwhile, in relation to capacity, some better-resourced local authorities have greater data collection and analysis capabilities than others. There is also the issue of instituting standard data-sharing processes and promoting the use of innovative data sources, which may entail significant levels of training on geospatial analysis and the use of new kinds of software. In this regard, ONS has sought to build not only its own capacity, but also central and local government capabilities more widely.

In an effort to meet local user needs and following a lengthy process of consultation with statistics professionals and departmental and political leaders at local authority level across the country, the Government Statistical Service (GSS) in the UK adopted a subnational data strategy that aims to:

- Produce more timely, granular, and harmonised sub-national statistics, the goal being to:
 - Think subnational by default – the national statistical service should not be producing only national statistics, if there is a way to make them subnational;
 - Investigate alternative data sources and incorporate them ensuring quality, accuracy and confidentiality;
 - Explore existing and new statistical methodologies, such as in relation to producing estimates for small areas and controlling disclosure;
 - Promote harmonisation of subnational data to make them more comparable, consistent and coherent; and
 - Enable robust, reliable disaggregation and intersectional analysis at a range of geographical levels. For example, the data may be used to explore the demographics of a particular area in terms of its residents, workers and vulnerable population, exploring the points of intersection among these groups.
- Build capability and capacity for subnational statistics and analysis, the goal being to:
 - Build the capacity needed to use and exploit geospatial data and methods;
 - Improve the way in which subnational data and metadata are shared across the country, including with the researchers.; and
 - Improve the discourse on the methodologies used to produce subnational statistics.
- Improve the dissemination of subnational statistics, the goal being to:
 - Ensure stakeholders are informed about new and updated subnational outputs so they are easy to locate;
 - Make subnational statistics accessible to a wide range of users; and

⁶ This section is based on presentations made by Emma Hickman, Deputy Director, Subnational Statistics and Analysis Division, United Kingdom (UK) Office for National Statistics (ONS); and Jim Hawkins, Subnational Development Analyst, ONS.

- Guide users on how to use subnational statistics and how to communicate their quality. For example, risks may be taken when innovating in the production of statistics, which can mean that the quality of such data may be more open to question. The users of such data should be made aware of this, so that they, in turn, should take care in how they deploy and present this information.

In this context a key aspect of the UK's national strategy has been to produce data that is accessible and meaningful to local stakeholders. To this end, it has developed an "Explore National statistics" (ESS) digital service that allows users, from expert analysts and policy influencers to inquiring citizens, to find out more about their local areas. The service, which has been developed through a series of "discovery", alpha and beta phases, offers users access to data on local areas and provides tools enabling them to visualise, compare and download this information.

The development of the service was undertaken iteratively, researching user needs and producing a series of products which were tested in terms of whether they were meeting those needs. The philosophy was one of continuous improvement, which nevertheless allowed for the creation of an interim product which, while imperfect, was of immediate use. Having gained a broad understanding of user needs and operating with the goal of implementing government policy on devolving some aspects of governance, the first stage was the establishment of a minimum viable product, which was launched alongside the government's Levelling Up White Paper in February 2022.

This alpha version of the digital service provided 36 indicators in its interactive version, covering all the metrics that had been identified as important during the preliminary research and testing phase and which were available at that time. A further eight indicators, which were more limited in geographical scope and were less amenable to useful comparison, were covered in the downloadable data on offer.

The data provided by the service at this time met identified user needs in relation to:

- Finding datasets about a particular, specified area (as well as the metadata and information on the methodology underpinning the production of these datasets) with the goal of informing decisions being made by these users at work;
- Visualising the data on a map or in the form of charts so that it could be used for further analysis or dissemination;
- Comparing data by different levels of granularity and by geography between and among relevant areas and regions, and against national benchmarks; and
- Downloading data in relation to the various indicators, allowing users with appropriate analytical capacity to build their own dashboards and deploy application programming interface (API) software to automate the incorporation of the data into their systems.

In terms of its quality, the data offered through the service comprised official statistics that had already been quality assured and data from other sources which had been verified in terms of the code of practice for government statisticians which sets standards in relation to trustworthiness, quality, and value.

Once the usefulness of the alpha version of the ESS had been verified through talks with local authorities, work commenced on producing a beta version. This entailed deploying tools that had been developed for the country's 2021 census, and incorporating a number of other available economic datasets – which was deemed a cost-effective approach.

However, the deployment of the census produced a number of challenges. The census tools offered little in terms of relevant changes over time and were also limited in terms of the areas that they covered. Accordingly, the beta version of ESS introduced a “changes over time” tool; a tool allowing users to build custom profiles for specific areas; and a “subnational indicators explorer” allowing users to explore areas with reference to different kinds of data, for example, through a local authority-level health index. The process of shaping the ESS product was influenced throughout by user feedback, including in relation the kind and detail of data required; the comparability of the data; the ways in which the data should be visualised; and the data’s reliability in terms of, for example, confidence intervals.

Significant efforts have been made to strengthen the metadata framework that underpins the digital service, finding ways of harmonising data from across government that has been produced on different topics in different ways by different departments. All the data has been cleaned and sorted by standard geography codes, revising obsolete codes to ensure that time series analysis can be undertaken. In addition, efforts have been made to ensure that the data pipeline can be updated readily and relatively quickly, making it more efficient and facilitating its management by other teams. To this end, modularised python scripts have been deployed.

The new beta version of ESS provides information to decision makers in a way that they can easily understand without a need for great analytical capability. In relation to finding, visualising, comparing and downloading data, the service enables users to:

- Find:
 - Search by area name;
 - Search by postcode;
 - Search by dataset;
 - Search by topic;
 - Find metadata and methodology; and
 - Access ESS via the ONS website.
- Visualise:
 - Data on a chart (by single area);
 - Data on a chart (by multiple areas);
 - Data on a map;
 - Data over time; and
 - The uncertainty of the data.
- Compare:
 - Against national average;
 - Against other areas;
 - Against statistically similar areas; and
 - Against other levels of geography.
- Download:
 - A full dataset;
 - Selected areas in a dataset;
 - Selected periods in a dataset;
 - Selected areas across multiple datasets (this is still a work in progress);
 - Automated data retrieval; and
 - Maps and charts, which may be embedded.

In addition, an analytical advisory service coordinated by regional teams has been established to help local government take full advantage of ESS. This service, which is called “ONS Local”:

- Provides analytical advice, supports analytical projects, and collects user priorities to inform ONS's analytical plans;
- Supports users in their efforts to navigate the subnational data landscape and access data platforms, and connects them to expert advice at ONS;
- Hosts regular events based on analytical themes, shares knowledge, seeks to engage users in existing networks, and deploys ONS's political/analytical knowledge to facilitate evidence-based decision making;
- Leverages ONS's position in central government to understand, identify, and align priorities for both local and central decision-makers, while capturing user needs; and
- Supports, informs, and advises local colleagues on data-provision and analysis opportunities, with the aim of continuously developing the offering.

Under the ONS Local scheme, for example, a number of workshops have been held in response to identified user needs, providing training on the use of APIs and Microsoft Power BI dashboards to allow automation of processes. A second series of workshops attended by hundreds of participants are now being held on this topic, drawing on existing expertise among the national community of statisticians and data analysts.

Questions⁷

ONS offers data in a standard template which has the virtue of offering uniform, consistent and comparable data. So, how do you address local authority needs and demands for their own context-specific data provision and analysis?

ONS local teams which network with local authorities provide support for more localised analysis and the development of bespoke local services. For example, after two local authorities said that they wanted to explore the issue of digital exclusion in relation to service-delivery, a one-off *ad hoc* project was launched to map broadband coverage in those areas. So, those local authorities now have the information that they require to understand the extent to which they can provide a lot of their services via the web, or whether they should still be deploying other channels for delivering these services.

In the South African context, governance is quite centralised and top-down, which can act as a disincentive to the implementation of a plan that must be operationalised at the local level, and which prioritises the dissemination of data that meet local needs. How may this problem be addressed?

In the UK, the new ONS initiative has depended on an acknowledgement within the national government that local government tends to know a lot more about what is happening in their areas than central government does. In other words, it is understood that there is an obligation on the part of central government to listen to local concerns and needs as expressed by local stakeholders; and to adapt the national infrastructure which central government has established and manages to address those concerns and needs. In large part, this perspective is the result of a push by central government to promote devolution in line with its Levelling Up White Paper. So, ONS is working alongside the Office for Local Government on the devolution project, which necessitates not only data-sharing in support of devolved evidence-based decision-making, but also effective accountability at the right levels within the new governance structure.

It is one thing to develop a prototype for using data more effectively in support of decision-making, and quite another to ensure that people use it as part of their work. In addition, under the agile-delivery approach, there is the issue of communicating that the prototype is a work in progress and that some problems may be encountered but that this is part of the process. How did ONS address these challenges?

⁷ This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

In rolling out ESS, ONS convened a series of webinars called “ONS Presents”. These introduced key local government stakeholders to new datasets, statistics or aspects of the interface as they were incorporated, with the aim of showing them how the new offering may be used to empower them further in their work. In addition, there was a strategy of trailing prototypes in advance of their launch to generate excitement and research their usability. So, for example, the beta version of ESS was tested by 200 local users, 97% of whom said they were likely or very likely to use the tool.

Central government stakeholders also were informed about and engaged in the roll-out of the programme. Colleagues from the Department for Levelling Up, Housing and Communities, which funded a lot of the work, His Majesty's Treasury, and the Cabinet Office were involved. Meanwhile, at a more strategic level, a senior subnational data group was established and chaired by the National Statistician to discuss evidence gaps across government and how these should be and have been filled by the various government departments. This forum was also used to showcase tools being produced by ONS that would be of use in supporting the work of these departments and their ministers so that they can be more effective.

In addition, ONS has shared the code for ESS, making the tool open-source, with the aim of promoting its reuse and longevity. Meanwhile, the agile-delivery ethos, which has already been widely deployed by a number of projects established as part of the national government’s digital service, is promoted via a label at the top of the tool indicating that it is a prototype, as well as a link inviting feedback.

4. MAKING LOCAL DATA ACCESSIBLE – SOUTH AFRICA’S APPROACH⁸

Statistics South Africa does not have control over the production of administrative statistics, which can lead to significant challenges when it seeks to produce data based on such statistics. For example, in compiling South Africa’s first “Migration Profile Report”, Stats SA reviewed a number of data sources relating to migration produced by various departments. The departments had deployed a wide range of collection and categorisation processes in producing these sources, which reflected their different statistical capacities and led to great inconsistency and problems of definition among the data on offer.

In addition, there are significant impediments to data sharing in the South African context; and there is the great challenge of capacitating municipal officials and decision-makers so that they can interpret and analyse data effectively, translating it into knowledge that can be used to produce implementable outcomes that benefit local people. At the same time, it is clear that there is room for improvement in the integration of data from various places, particularly administrative records, with the aim of producing useful synthetic data sources.

For its part, Stats SA produces a number of statistical series that provide data at local levels even as there are limitations on the disaggregation of the underlying data at this level. The datasets are produced using empirical sources, but also deploying administrative records.

So, for example, mid-year population estimates are based on empirical data collected through the census, but updated in the intercensal period using administrative records, including those relating to births and deaths; internal and international migration; numbers of pupils via the Department of Basic Education’s (DBE’s) Learner Unit Record Information and Tracking System (LURITS); and voter registration from the Independent Electoral Commission, as well as information held by the Department of Health. The lowest level of disaggregation for this population data is the local level. In addition, this data is deployed to produce short-term (five-year), medium-term (10-year) and long-term (35+-year) projections at the local, district/metro and provincial/national levels respectively.

Stats SA also produces an annual general household survey, which looks at household dynamics and education among other issues; an annual domestic tourism survey; and an annual Governance Public Safety and Justice survey – all of which offer data disaggregated to the metro level. In addition, it produces a quarterly labour force survey which is disaggregated to the metro level.

Stats SA also produces a number of surveys on behalf of other government departments, including: a National Household Travel Survey, which is undertaken every five years and provides data disaggregated to the district or transport-zone levels; and a Demographic and Health Survey, which deploys a standard international questionnaire and provides data disaggregated to the provincial level.

In addition, Stats SA collects administrative data from a range of government sources, producing reports from this information on:

- Tourism and migration. Based on data from the Department of Home Affairs about the number of people crossing the border either as visitors or for another purpose, this information is available at national level only.
- Marriages and divorces. Based on marriage records from Home Affairs and divorces data from the Department of Justice, this annual report provides data disaggregated to the provincial level.
- Mortality and causes of death. Based on the death notification forms issued by Home Affairs, this data, which is reported annually and is disaggregated to district level, raises the issue of protecting the confidentiality of those are no longer alive and so cannot speak for themselves.

⁸ This section is based on a presentation made by Statistics South Africa

- Births: Based on registration data held by Home Affairs, this annual report disaggregates annual data to the provincial level.

Stats SA also conducts an intercensal Community Survey every five years which offers data disaggregated to the local level. The present plan is to deploy the next community survey as part of a new Continuous Population Survey (CPS), which would entail the implementation of a rolling census undertaken on a three-yearly basis.

Questions⁹

Given the importance of understanding levels of migration in producing policy, are there any plans to disaggregate the data produced as part of the Migration Profile Report beyond the national level, and will data on internal migration be included?

Following the online publication of the Migration Profile Report, a number of user engagement sessions will be convened and the feedback from these will determine the level of data disaggregation. Given that this is the first such report, it has been part of a learning exercise, producing an understanding of the relevant stakeholders and what information is available and where – all of which provides a basis for improvement. In particular, future such reports may seek to integrate data from provinces, metros and municipalities on how migrants are engaging with the state in terms of service provision and whether they may be moving as a result of their experiences of such provision. So, there will be disaggregation of the data at the provincial level, although not so much at the local level; and it will cover internal as well as international migration.

To what extent will the collection and analysis of data, which may be provided in collaboration with private-sector actors, affect national economic policy-making?

The national statistics system engages government entities in relation to the coordination and uniformity of data collection. In this context, Stats SA's sole mandate is to collect data. In this respect, it may monitor whether changes in policy should lead to changes in the indicators used to collect and analyse data, but it does not involve itself in shaping the policy-driven mandates of the various departments. Its engagement is limited to identifying what kind of data is available; and what kind of methodological approaches are being employed, so that it can fulfil its own mandate of collecting and sharing relevant statistics.

The general household count and the indigent count which are undertaken for the General Household Survey are insufficient as an evidence-base for municipal decision-making. In this respect, is Stats SA looking to provide data of greater local use?

The General Household Survey is weighted using household estimates every June and enables measurement of the population in terms of numbers, gender and age, as well as a general sense of household dynamics, which is the key aim. However, the sample size may be too small to enable a focus on a particular sub-sector of households. Such a focus may be better supported through analysis of administrative records held by the metros and local municipalities. At the same time, Stats SA appreciates the value of investing in administrative data with the goal of supporting evidence-based decision-making in answer to South African needs at the local level.

Does Stats SA look at women-owned businesses, for example, in the property sector?

At present, Stats SA does not produce business surveys. It issues data on volumes of production in the manufacturing sector on a monthly basis, and also produces reports on mining activity. The goal going forward it to produce statistics for all the industrial sectors.

⁹ This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

Could a question on place of work be included in the national census in support of efforts to produce spatial economic data at the local level?

There used to be a question on this in the census. If there is a good rationale for including it again in response to feedback from stakeholders, it may be reinstated. The question may have been removed on the basis that it was considered of little use to stakeholders at that time. Also, this is not a question that is generally included in censuses internationally, unless as part of a travel or transport survey.

5. GAPS IN THE LOCAL ECONOMIC DATA SYSTEM¹⁰

Participants at the workshop split into four breakaway groups which identified:

- A range of important data sources;
- A number of challenges and gaps in the present provision of data in South Africa;
- A range of opportunities for the provision of useful data;
- A range of opportunities for capacity-building in the field of data provision and analysis; and
- A number of lessons that may be learned from the UK's experience of providing local spatial economic data.

Important data sources

- SEAD-SA;
- The National Income Dynamics Study (NIDS) produced by the Department of Planning, Monitoring and Evaluation (DPME) with the Southern Africa Labour and Development Research Unit at the University of Cape Town (UCT);
- Quality of life data provided by the Gauteng City-Region Observatory;
- Other think-tanks and researchers;
- National census;
- Financial and economic data provided by Stats SA in the form of the Quarterly Financial Statistics (QFS) and Quarterly Economic Statistics (QES);
- Workforce data provided by Stats SA's Quarterly Labour Force Survey (QLFS), including data on those not in employment, education or training (NEETs);
- Manufacturing data provided by Stats SA;
- Companies and Intellectual Property Commission (CIPC) data on firms;
- Cadastral data from land surveyors which may be included in property deeds;
- National Treasury data on infrastructure projects;
- Road, rail and ports data from the South African National Roads Agency Limited (SANRAL), the Passenger Rail Agency of South Africa (PRASA), Transnet and the Department of Transport;
- Gross domestic product (GDP) data produced at national, provincial and municipal levels;
- Foreign direct investment (FDI) data;
- National trade data;
- Education and training data provided by national and provincial government departments, including the Department of Basic Education and the Department of Higher Education and Training (DHET), as well as the Sectoral Education and Training Authorities (SETAs);
- National and provincial public health data;
- Service-delivery infrastructure data on roads, water reticulation and electricity provision which may be provided by municipal departments responsible for geographic information system (GIS) data;
- Municipal data on local business-licence applications;
- Municipal financial statement and budgets;
- Municipal building planning permissions;
- Municipal property valuations;
- Municipal recreation data on parks and pools;
- Municipal transport data on bus routes and activity;
- Tourism data on hotel occupancies and movement of people internationally and inter-provincially;

¹⁰ This section is based on feedback from four breakaway groups on this topic.

- Cellphone data from the main telecommunications firms (over 60% of the population use cellphones); and
- Data from private data providers, such as Quantec.

Challenges and gaps in data provision

- There are issues around the credibility of the local data produced by national providers which can be difficult to triangulate;
- By comparison with the UK, there is less localised data produced centrally, including by Stats SA – so, municipalities have a relatively greater role to play in developing their own data, including through surveys;
- Data can be produced at the municipal level with the main intention of ensuring compliance with national directives and with little regard for how useful it may be beyond that;
- There is a lot of useful information in municipal integrated development plans (IDPs) and spatial development frameworks but it can be difficult to extract and analyse the raw data underpinning these documents;
- In tracking graduate unemployment, there is a need for synergy between the data produced in the public and private sectors monitoring whether workforce skill needs are being met, and the data being produced by universities on graduate employment outcomes; and
- It is important to ensure that data provision protects personal information in line with the Protection of Personal Information (POPI) Act of 2013.

Opportunities for the provision of useful data

- Unemployment insurance fund (UIF) and social grant data may be more effectively mined for local data;
- Property tax data held by municipalities may be integrated into spatial economic data sets;
- Local household surveys, such as that on garages on parking which was conducted in Cape Town, may be expanded and adapted to produce useful spatial data;
- Data on planning permissions, which includes geospatial information, may usefully be integrated into other data sets;
- The drive to provide data in support of access to market for small, medium and micro enterprises is hampered by a lack of information on the informal market, as well as a lack of agreement of how informality may be defined. In this regard, a useful first step may be to define informality by scale of business as including those who are identified by SARS as sole traders on the basis of their individual identity documents. On this basis, useful data about informal markets may be produced;
- Small businesses can come into being, flourish and die quite rapidly. Data on changes in their rates payments over time – for example, when a residential property is re-registered as a commercial property – may be used to track the life-cycle of such businesses as a measure of economic activity;
- Useful data can be held in silos. For example, service delivery infrastructure which can have a major impact on economic activity may sit separately within GIS units/department. Such data should be integrated more widely to foster a better common understanding of local conditions;
- Data produced by researchers could be leveraged more effectively by local policy- and decision-makers;
- Perhaps municipalities could pay for national data collection to be expanded in their areas; and
- Local surveys may be designed to appeal to a wider sample by leveraging social media and deploying icons to make them more user-friendly. For example, emoticons may be deployed to measure levels of happiness with a city.

Capacity-building opportunities

- Data scientists and local officials should be upskilled to strengthen this sector; and
- Stats SA may have a role to play in upskilling statisticians within local government in a bid to ensure harmonisation in the definitions deployed to shape data, and how such data is collected and organised.

Lessons that may be learned from the UK's experience

- The UK has adopted a programme of deliberately identifying and trying to meet the data needs of local authority users – South Africa could adopt a similarly responsive ethos in its production of data;
- The UK's use of live administrative data provides a model for the production of regularly updated data;
- The UK's experience shows the importance of reviewing existing policies, including in terms of their implementation, instead of just adding more policies. South Africa already has excellent policies; the trick is to find effective ways of implementing these – perhaps through use of an agile methodology which allows for the continuous development of prototypes informed by user feedback; and
- In this regard, South Africa's service delivery problem may be most effectively addressed by considering user experiences and instituting mechanisms to monitor these in order to measure the effectiveness of provisional solutions that are implemented.

6. UK's approach to using administrative and 'big' data¹¹

ONS uses regular and live feeds of administrative data as real-time economic indicators for the weekly release of information on: consumer behaviour; business and workforce activity; energy use and activity in the housing market; and international transport connectivity.

The weekly release on consumer behaviour features real-time economic indicators derived from:

- Data on spending from a national credit card transaction database, which provides a picture of spending and consumer hardship, for example, in terms of failed debit transactions;
- Fuel purchase data provided by the big retailers; and
- Transaction data from food retailers, such as Pret a Manger (which also helps to provide detail on footfall at train and bus stations that was used under Covid-19 to track the impacts of the pandemic).

The weekly release on business and workforce activity features real-time economic indicators derived from:

- Online job adverts which are tracked by a company called Adzuna;
- Companies House data on newly incorporated and dissolved businesses, which provides information on where new businesses have opened and when they have closed, as well as a picture of potential redundancies; and
- Data on value-added tax (VAT), or sales tax, for which all firms trading in the UK have to register.

The weekly release on energy use and activity in the housing market features real-time economic indicators derived from:

- Data on energy performance certificates, which are required for every home sale and indicate the house's energy efficiency. These provide a picture of the quality of the housing stock;
- Rental analytics data, which provides a picture of the affordability of rental accommodation;
- Gas and energy price data from the national grid; and
- National gas transmission data.

¹¹ This section is based on a presentation made by Jim Hawkins, Subnational Development Analyst, ONS.

The weekly release on international transport connectivity features real-time economic indicators derived from:

- Satellite automatic identification system data from a company called ExactEarth, which provides information on total ship visits, including by tankers and cargo vessels;
- Traffic camera data which provides information about road usage; and
- EuroControl data which provides civil aviation information, including on the number of flights taking off from and landing in the UK.

ONS is also undertaking a project to enable the production of data on gross disposable household income (GDHI) at a granular level by lower super output (LSO) area, which is the smallest population-based geographical unit published by UK statisticians. Such data is useful for assessing the impacts of economic plans, such as in relation to infrastructure, at the neighbourhood level – for example, whether the construction of a new rail line or train station has had an impact on GDHI in the local LSOs. This kind of data is of particular interest to local authority planners, as well as business and other stakeholders.

GDHI comprises the amount of money that individuals in the household sector have available for spending or saving, after deducting expenditure associated with income and property ownership, such as taxes; employee social contributions; and provision for life insurance. It is calculated gross of any deductions for capital consumption, which is the decline in the value of fixed assets due to normal physical deterioration and obsolescence in the household sector.

A range of administrative data sources are deployed to produce GDHI at the level of LSO units (which are administrative units within local authority structures) with the proviso that the data can actually be disaggregated to this level. Where it has been found that one component of GDHI which may be available at a granular level correlates with others which may not be available at the level, then the former will be used as a proxy for the latter. In addition, the production of GDHI at LSO level is being undertaken iteratively with the aim of continuously improving the accuracy and usefulness of the indicator.

The components of GDHI include employment income and income received from the property as provided by data from the tax office; data on dwelling stock and housing rates; and data on social benefits received, which is available from the relevant government departments. In addition, the calculation of GDHI includes an “operating surplus” component which represents income (imputed rent) derived as a result of the value of the property, which is based on local authority data on the dwelling stock.

ONS also uses large administrative datasets to produce an Interdepartmental Business Register (IDBR), which is published on a quarterly basis. UK firms that are trading have to register for VAT; and all employers have to register for pay-as-you-earn (PAYE), which is the system for deducting tax from employees’ wages and salaries. This data provides a picture of the number of active businesses; their employees; their turnover; and how their local units or operating sites are distributed across the country. In addition, all firms must register their establishment at Companies House, which provides a picture of businesses that have dissolved, and where new businesses are being established; this data also provides an up-to-date baseline for the other data sets, ensuring they are current and accurate. The register sorts the information by enterprise; enterprise group owned by a single entity; and local units, which are the operating sites of businesses. This offers the basis for a granular analysis of business activity.

ONS has also become involved in efforts to promote greater statistical coherence in data held by different department on public-sector expenditure. As part of efforts to assess levels of public expenditure on research and development (R&D) at the regional level, ONS uncovered a startling lack of coherence in the ways in which data on this was being collected and produced by a number of large government departments, including the Ministry

of Defence; the Department for Health and Social Care; and the Department for Business, Energy and Industrial Strategy; as well as by “arm's length bodies” under the aegis of departments.

In terms of granularity, some of the R&D data included postcodes and described in detail where the funding had been spent and by whom. Other data, however, only listed the addresses of the parent organisation headquarters with no indication of where the research and development had been performed; or failed to disaggregate R&D expenditure from other expenditure on the same programme. Some of the data would have high granularity, while other data would be quite incomplete. Some expenditure data would include overheads and staffing costs, some would only include gross expenditure on R&D. In general, the discrepancies were due to the range of terms and methods used to record the procurement process, as well as the different rationales for collating the data – that is, the ways in the data were supposed to meet the specific needs and purposes of the particular team overseeing the project.

However, such discrepancies, for example, in relation to location data and the granularity of the data being produced, can impede the establishment of an integrated data system and the use of such data for statistical purposes. In response, ONS has undertaken an interdepartmental programme to develop, improve and standardise public expenditure data, issuing guidance that departments should appoint coordinators who are responsible for ensuring the standardisation of the data being produced by teams, while also providing support so that the data continues to be produced in ways that address the particular team's or department's needs and purposes.

Discussion and questions¹²

In response to the presentation, a number of challenges in relation to the availability and reliability of administrative data sources in South Africa were noted. Key national administrative data sources include:

- The national population register, although many South Africans do not have identity papers and birth certificates;
- The voter's roll held by the Independent Electoral Commission (IEC), although many residents are not registered to vote;
- Data on those in education provided by the Higher Education Management Information System (HEMIS) and Department of Basic Education data;
- Tax data provided by SARS; and
- National expenditure data offered by the National Treasury, although there are challenges around the real-time nature of this information and around its accuracy, given that it is based on inputs provided by municipalities which may classify expenditure improperly.

Meanwhile, the provision of data on business activity is hampered by financial constraints, with current surveys and data only allowing analysis at enterprise level only rather than plant/outlet level.

Does the UK incentivise stakeholders to give information freely? In South Africa, there are challenge around accessing privately sourced data – for example, Vodacom charges for its data.

Under a Digital Economy Act, which was introduced in 2017, government departments are required, where they can and where it is feasible, to share data with each other. In support of such data-sharing, efforts have been undertaken to identify areas where different departments have common goals and then to promote an understanding that sharing data in pursuit of these is in their interests.

¹² This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

The Act does not just apply to departments, it actually applies to any data providers in the UK. In this respect, however, ONS does not adopt an enforcement approach in seeking to access data from the private sector. Rather it engages key firms in a conversation about how ONS may support them in, for example, quality-assuring their data – and then, on the basis of the relationship that has been established, identifies which data sets held by the firm may be of wider use, bearing in mind the need to protect commercially sensitive information.

For example, having established strong relationships with Visa and Barclays bank over a period of years, ONS was able to persuade other institutions in the banking sector to provide comprehensive financial activity information as part of government efforts to track and understand the economic impacts of the Covid-19 pandemic. Subsequently, this data has continued to be provided and is produced by ONS in the form of a weekly feed used to understand economic impacts.

ONS does not pay commercial prices for such data in part because of the value that it can offer as a partner, and in part because there is a willingness among private-sector actors to cooperate and collaborate with the government for strategic reasons. In fact, ONS tends to just cover the data-processing costs. In addition, it tries to negotiate for broad rather than one-off usage in the contracts that it establishes with private-sector data providers.

How does the UK maintain its business register? Are businesses that close eliminated from the register? And what if they return?

The national business registry is updated on a quarterly basis and, in cases where a firm or business has closed, a note is made on the entries recording the date of death. However, the data on the company is not removed, so it is possible to track the churn in the number and kind of businesses being established and expiring.

To what extent does ONS engage with and encourage researchers in universities and in the private sector to use the data that it produces in order to add to the sum of knowledge in society?

Much of the data produced by ONS is made available via a secure research service – so, researchers apply to access the data and they are checked on security grounds and in relation to how they are planning to use the data, which cannot be for commercial purposes. They may then be granted access to the data via ONS's secure data platform. In addition, ONS collaborates closely with academics in the production of its methodologies to ensure their quality and effectiveness. To this end, it has a long-term partnership with the Economic Statistics Centre for Excellence and has established a network of academics and statisticians in support of its work.

7. ADMINISTRATIVE DATA IN SOUTH AFRICA

7.1 Overview of the UNU-WIDER Secure Data Facility¹³

The aim of the SA-TIED programme is to produce data and research that can inform policy-making. In this context, many of the research projects undertaken through the programme at the Secure Data Facility in the National Treasury were designed in collaboration with policymakers, which means that the outputs will be seen by policymakers and may well inform their efforts.

Broadly, the purpose of the National Treasury secure data facility is to:

- Make anonymised administrative tax data available for research purposes;
- Foster policy relevant research; and
- Act as a strategic meeting place where leading local and international researchers and National Treasury employees can come together around the “watercooler”, as well as at seminars and workshops, to discuss and address administrative-data and policy-research questions and needs.

Since its establishment in 2014, when it only housed three computers and was used by a handful of data scientists and researchers, the facility has greatly expanded its capacity as demand for access to the data sets that it holds has grown among researchers and policymakers. At present, those using the facility need to book a month in advance and the facility is planning to install more hardware to keep pace with the growing demand.

The anonymised data that is hosted at the SDF derives from data shared with the National Treasury by SARS. This data is high-quality and quite comprehensive, although it does not cover the informal sector. It enables longitudinal tracking over extended periods and offers a much cheaper form of data provision than traditional methods, such as surveys. On the downside, it can be quite messy and is generally unsupported by significant documentation. In addition, the data was not created for research purposes.

At present, the SA-TIED programme is focussing on a number of research topics in its deployment of the data at the facility. These relate to:

- Climate change and energy;
- Automation, the labour market and productivity;
- Evaluating policy reforms;
- Firm audits and tax gaps;
- Market power and competition policy;
- Income and wealth inequality
- Education, skills mismatch and frictions in the labour market;
- Market access and wage disparities; and
- The gender wage gap.

Researchers working on some of these topics use outside data in addition to that available at the Secure Data Facility. In this regard, one of SA-TIED’s strategic goals is to enable access to other administrative data sources. In 2018, there were 25 research projects under SA-TIED, with the number increasing to 72 in 2020, before dropping off as a result of the Covid-19 pandemic, although the number of projects has subsequently risen again.

The SDF is not connected to the internet in order to ensure the security of the data stored there. Whenever researchers want to take data away, they have to send the request to a member of the SA-TIED team who then

¹³ This section is based on a presentation made by Adaiah Lilenstein, Data Lab Manager, United Nations University World Institute for Development Economics Research (UNU-WIDER).

assesses whether the data in question can be removed or whether it is too sensitive and needs to be manipulated further before it can be taken.

The data pipeline itself starts with the taxpayers and firms that fill out tax forms that go to an SQL (structure query language) data warehouse where SARS then collects, cleans and manipulates data before providing it on a physical hard drive to National Treasury, which is then plugged into another SQL data warehouse. The data is then extracted, cleaned, checked and separated into firm-level and individual-level panels that can be shared with researchers.

The main administrative tax datasets available at the SDF relate to: corporate income tax; payroll tax certificates; value added tax; customs (import and exports); personal income tax, excise tax; dividend data; common reporting standards; employment reconciliations; and labour brokers. New datasets that may be useful for research are always being requested, although these cannot always be provided by SARS in a publicly available form. An employer/employee panel dataset has been established at the SDF using the data from corporate income tax, payroll tax certificates, VAT and customs; and an individual panel has been created using the data from payroll certificates and personal income tax.

There are three ways of accessing the data:

- Via the SA-TIED programme, which commissions and funds research on a number of priority topics;
- Via the National Treasury, if it deems the research topic policy relevant, feasible and novel, although there is no funding through the programme for such studies; and
- Via the SEAD-SA online spatial tax data portal, which provides a host of granular and aggregated data on a number of economic indicators that have been pulled from the tax data.

7.2 Vision for an integrated administration database for South Africa¹⁴

It has been proposed that, building on a number of existing initiatives that seek to deploy and integrate administrative data in support of public- and private-sector policy- and decision-making, a more integrated approach to the production and development of such data may be adopted – and a South African “integrated data lake” (IDL) may be created.

The idea builds on three current initiatives:

- The establishment of the Secure Data Facility at National Treasury, which has leveraged data supplied by SARS to provide information that researchers can use to help build an evidence base to inform policymaking;
- The establishment of an Integrated Business and Individual Register (IBIR) by SARS which deploys a range of different administrative datasets to which the service has access (notwithstanding some resistance and delays) as a means of supporting revenue collection processes and improving auditing of financial returns to ensure compliance with tax requirements. This register leverages CIPC data; vehicle registration data; information from the National Population Register; and a number of privately held datasets, such as, for example, on private airplane registrations. It should be noted that the use of the data accessed by SARS for the IBIR is restricted; and
- The Spatial Tax Panel project which has sought to provide anonymised datasets from SARS data aggregated into a range of spatial and temporal units and correlated to a range of indicators with the aim of providing a sense of what might be happening economically on the ground in the country’s metros. In the absence of any other such granular data, this information has been used to produce a range of valuable insights and provide answers to questions that were previously unanswerable.

The aim is to incorporate the data from these sources into the IDL alongside data from a range of other data sources, which may include:

- Grants data from the South African Security Agency (SASSA) and the Department of Social Development (DSD);
- Survey data and statistical products from Stats SA;
- Data on UIF payments and claims from the Department of Employment and Labour, which would provide information on people moving in and out of employment;
- Data from the National Population Register, which would provide an understanding of the changing patterns of intergenerational relationships;
- Data on schools and universities, and pupils and students from the Department of Basic Education and the Department of Higher Education and Training;
- Deeds, cadastral and other spatial data from the Department of Agriculture, Land Reform and Rural Development (DALRDD);
- Patient data from provincial and national health departments; and
- Data from a range of private sources such as credit bureaux, banks and aviation and shipping authorities.

The establishment of an IDL would depend on creating a culture of trust so that different government departments and agencies would be prepared to share the datasets that they hold with each other. Such a culture may be fostered by creating and implementing agreed standardised processes for sharing data. Such processes should ensure the appointment of individuals within departments who would be responsible for ensuring the production,

¹⁴ This section is based on a presentation made by Andrew Nell, Consultant, Spatial Economic Activity Data – South Africa (SEAD-SA).

collection and sharing of appropriate data, with succession-planning mechanisms in place to ensure that these positions are always filled.

The intention is to send the data to a secure offline environment where it will be curated and integrated by a small team of professionals, who will, as necessary, provide feedback to the original owners and producers of the data owners about any issues that they may find; and who will find ways of aggregating the data, including sensitive data, so that it can be of use to policy- and decision-makers, researchers and the general public. In this regard, it is hoped that those working at the integrated data lake will be provided exemption under the POPI Act, allowing them to access the data in its raw, de-identified form so that they can find ways of integrating and developing anonymised datasets which are as useful as possible.

The intention then is to share the data in a number of ways, including through automated processes that would make live data from a number of the administrative sources available via APIs. This data, which would be supplied in line with the safety protocols of the various providers as part of a raft of IDL digital services, would then be incorporated in the data held at the secure offline environment, and also, as appropriate, via an online portal.

Having been sorted and processed at the offline IDL facility, the data would be made available through three main channels:

- The Secure Data Facility at the National Treasury, as well as another SDF, which could be housed at Stats SA. Deploying strict safety protocols on access to the data which may be made available remotely, the aim would be to provide a wide range of de-identified datasets, as well as a number of new merged data panels for researchers, and public- and private-sector stakeholders as appropriate. For example, such a service may offer panels that merge tax and education data, providing an understanding of education outcomes; or tax and grants/UIF data, providing an understanding of people moving in and out of employment.
- A South African IBIR. Such a service, which would be made available to SARS, would build on the IBIR model already established by the national tax collector. The aim would be to expand the regulatory and compliance components of the register, engaging with the various regulatory bodies and assessing what datasets they might have a mandate to access, and then trying to find a way of accessing these. This may be made possible through the IDL's drive to standardise and streamline data-sharing processes with a guarantee of security across government departments.
- An online portal. This would provide controlled access to a range of datasets that may be downloaded, with some users enjoying greater rights of access than others. The data provision would be supported by the provision of dashboards and data stories that would aim to make the data as accessible as possible. Provision would also include a catalogue of all the datasets that exist within the broader South African integrated data lake, including and beyond those directly available through the portal, with contact details for the owners of these data sets. The aim of this would be to provide a sense of what available data exists across government and also to allow users who so wish to try and source some of the immediately unavailable data through a standardised request process.

Questions¹⁵

The IDL seems to offer national horizontal integration of data but little in the way of vertical integration with and for local authorities – is this deliberate?

¹⁵ This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

The list of data sources in this presentation is not comprehensive. The data sources that are available also include municipal and provincial data sources.

What of the other stakeholders, such as those working in the taxi industry who don't fill in tax forms? In other words, what of the informal economy?

There is available taxi data which should be included in an IDL. That said, there are obvious gaps in data provision, for example, in relation to the informal sector. It is difficult to find ways of readily quantifying economic activity in the informal sector or to identify useful sources of data that may illuminate understanding. In fact, this is a key concern and topic in the discussion to establish an IDL.

Would the IDL be a physical facility? What kinds of similar facilities have other countries established?

In terms of the IT infrastructure, much may be learned from the kinds of infrastructure and models for such data integration and dissemination which have been established elsewhere, and how and whether these could work in South Africa.

Could the data available through this facility be used to help political parties?

In terms of general access, the data would be made available or not depending on its sensitivity, including to members of the general public and political parties. In relation to access by researchers and other stakeholders with a particular interest, South Africa may learn from the UK which assesses whether those seeking to access data are planning to use the information in a way which promotes the public good. This criterion offers a sound basis for whether or not to grant access, although there would need to be monitoring to ensure that those who have accessed data are not using it in pursuit of parochial interests at the expense of the public good.

What of the data quality in this model? Is there not a need to clean and quality-assure the data before it enters the lake – which would require great time and significant expertise?

A key aim of the IDL project would be to promote enforcement of existing data standards which may not be being followed at present, in part because of a lack of transparency among government departments. However, once there is an imperative to release data and the information is shared, transparency is created and a conversation may be started with those who are responsible for collecting and holding data – a conversation that may consider issues of compliance with relevant legislation and metadata standards. The idea is that those receiving the data at the IDL would offer feedback to the data owners and collectors that would enable them to fix problems with how the information is being collected and curated at source. In this way, the IDL may save itself time and money that would otherwise have to be spent cleaning the data at a relatively late stage in the process.

What is the best approach to promoting the sharing and use of administrative data?

In seeking to develop a community of practice, there is a tension between prioritising data collection and aggregation on the basis of relative availability or in response to user needs. One way of addressing this would be convene some focus groups, bringing people together to identify a number of key data priorities. In this regard, discussions with other countries about the present project and similar projects have indicated a need to strike a balance between casting the net wide to incorporate a range of datasets and identifying what is achievable, as well as the importance of building merged panels that can support research on areas of interest (for example, combining data from the population register and tax returns to promote an understanding of the changing patterns of intergenerational wealth). In this way, the beneficial impacts of greater sharing and use of data can be demonstrated. In addition, the UK experience has shown that efforts to advance the sharing and use of data can also be promoted through a tiered approach to access that offers relatively broad access for the public at the same time as allowing researchers more granular access. This kind of regime for access promotes the work by indicating and communicating the benefits that may be derived from accessible, usable data as widely as possible.

8. SEAD-SA SPATIAL TAX PANEL¹⁶

One of the major challenges South Africa faces as country is the great level of inequality among the population, which is expressed in terms of space and place, for example, in relation to the nature of one's neighbourhood. In this respect, Venezuelan-American economist Ricardo Hausmann recently said: "We found two broad classes of problems that undermine inclusive growth in the Rainbow Nation: collapsing state capacity and spatial exclusion." Accordingly, South African economists must address the issue of inequality in relation to economic activity on the ground in order to produce a properly functional national economy.

In this context, the economic importance of cities, which account for almost two thirds of all formal jobs, must also be emphasised. The country's cities need to work if the national economy is to work. The importance of the role played by cities relates not only to their relatively great size, but also to the advantages that they produce for firms and individuals as a result of their capacity to concentrate activity spatially. In this regard, notwithstanding the pressures exerted on creaking municipal infrastructures and housing provision by the high numbers of residents, these places offer better economic opportunities than are available in rural areas. For example, there are 31 formal jobs for every 100 urban residents, compared with only five formal jobs for every 100 rural residents. Clearly, cities are places where firms can thrive, be productive and produce jobs, indicating the importance that should be attached to them in national economic policy-making.

Against this background, the establishment of SEAD-SA was driven by a major data gap in terms of the information available to municipal economic planners and policy-makers; and a lack of answers to basic spatial-economic questions, such as: What is the GDP of Johannesburg and how has this changed over time? Which industries have created the most jobs in the metros, including jobs for the youth? Do firms in eThekweni compete with or complement firms in Gauteng when reaching to inland markets? Which areas within Cape Town are leading or lagging, and in which kinds of industry? What is the economic trajectory of central business districts (CBDs) compared with other urban nodes? How does job creation compare among metros, secondary cities and farming areas?

In an effort to address the lack of credible, granular economic data that may be used to answer such questions, SEAD-SA has sought to promote the use of administrative data, such as that available through tax returns submitted to SARS. The use of such data offers a number of advantages in relation to:

- Accuracy – it does not depend on respondent answers that may shaped by factors including a lack of recall and fear of stigma;
- Coverage and depth – it offers large data samples that allow for fine-grained analysis;
- Frequency – it enables longitudinal tracking over extended periods; and
- Cost – it can be produced at a fraction of the expense of surveys and fieldwork.

However, there are a number of challenges that must be faced in the production of such data, including in relation to:

- Usability – the preparation and cleaning of such data can be a time-intensive and thankless task;
- Sensitivity – the data must be properly de-identified for legal and ethical reasons, and there is a perpetual tension between ease of access and unmitigated risk in the use of such data; and
- Gaps – the administrative data was not compiled for larger statistical purposes, including that of facilitating spatial-economic understanding.

¹⁶ This section is based on a presentation made by Prof Justin Visagie, Senior Research Specialist, HSRC; and Associate Professor, UF; and by Andrew Nell, Consultant, SEAD-SA.

The administrative data promoted by SEAD-SA comprises a Spatial Tax Panel, which is derived from PAYE tax records held by SARS, which account for every formal employer/employee relationship across the country. This data includes a place-of-work address with a postal code that allows for disaggregation that can be visualised at a local level in municipalities, as well as a host of other useful information.

The data is cleaned (eradicating duplicates and imputing missing information) and anonymised at the Secure Data Facility at National Treasury, so information such as the names and addresses of the firms is not available through the panel. The aggregated and masked data is then made publicly available online at spatialtaxdata.org.za; while for those who are interested, the raw data that forms the basis of the panel and a document outlining how it has been processed are also available.

The data in the panel is organised in relation to space, time, kind of output and type of aggregation:

- **Spatial variables:** The data is available at national, municipal and postal code levels, as well as at a local level in the metros, where the data has disaggregated by equal-area hexagons.
- **Temporal variables:** The data is available by tax year, with a number of employment-related indicators available by month.
- **Output variables:** The data is available by number of employees and number of business establishments, which includes the branches of large firms. It also offers information on median income levels and Gini coefficients.
- **Aggregation variables:** Supplementary data available from the tax records enables the provision of further detail, including in relation to employees, establishments and median incomes in a particular area. So, for example, there is information on the sex, age and immigration status of workers; how much they earn; and the industry sectors and kinds of establishment in which they work. There is also information on the size of firms and business establishments (in terms of numbers of employees, turnover and current assets); whether the particular establishment is part of a bigger firm with other branches/offices; the numbers of businesses that have been established or have closed; and the sector in which the firm operates, including whether it is involved in import/export.

The geocoding algorithm that has been used to disaggregate the data by postal code into spatial hexagons may be used to incorporate further spatial information available at the sub-municipal level to strengthen the accuracy and improve the quality of the data being produced at this micro level.

Meanwhile, it is important to note the limitations around the current deployment of administrative data through the Spatial Tax Panel:

- It only provides information from firms and individuals that are registered tax payers, so cannot be used to describe economic activity in the informal sector;
- The disaggregation of data to hexagon level from postal codes introduces a significant margin of error, even as this micro information can be of great use;
- The data can, in a few instances, be skewed by the “head office effect”, which describes instances in which firms list all their employees as working at the head office rather than at the various workplaces managed by the company. Although this is a relatively small issue given that only a handful of large firms report to SARS in this way, further work is being undertaken to quantify its impact.
- Establishment-level data is only directly available from firms reporting using the IRP5 forms which ask for this information. Such data is unavailable for tax-exempt institutions and must be inferred for firms that report via other methods.
- There are data lags of up to two years for specific variables due to the (annual) frequency for the delivery of the SARS data to National Treasury.

- In order to protect the anonymity of individuals and firms, all data is masked when there are fewer than 10 individuals/firms in a specific aggregation, which means that there is significant data loss when seeking to analyse such granular information across multiple variables.

The datasets produced by the Spatial Tax Panel have been benchmarked against the Quarterly Labour Force Survey and Quarterly Financial Statistics produced by Sats SA as a way of assessing their accuracy, and it was found that there was a strong correlation, including at the municipal level, indicating their reliability.

In an effort to make sense of the information provided by the panel; produce new knowledge; and indicate the potential of the data for economic planning, the National Treasury and HSRC produced a report, titled *Cities Economic Outlook 2023*. This presents key policy-relevant insights about the economic performance of South African cities distilled from the spatialised tax data. In particular, the report explores a variety of themes including: the role of urbanisation in economic development; the specialisation of each metro economy; the uneven impacts of Covid-19; the relationship between cities and productivity; and the central role of cities in the South African labour market. It further analyses the relative economic performance of cities, for example, by comparing that of eThekweni with that of Cape Town, which has produced some counter-intuitive findings.

In addition, SEAD-SA's Spatial Tax Data Portal at spatialtaxdata.org.za offers a user-friendly web interface for visualising, exploring and downloading the spatial tax data in ways that can be tailored to the needs of a particular municipality. The portal includes:

- Dashboards: Users can choose a dashboard related to the municipality of interest. The dashboards are a series of curated charts or figures organised by theme which help users gain insights into what is distinctive about a municipality. The charts are organised into four themes: “overview”; “economic growth”; “industry diagnostic”; and “equitable economies”.
- A map explorer: The map explorer visualises spatial tax data in maps. Tax data can be explored either by comparing municipalities on a map of South Africa, or by comparing suburbs on a map of a metro. Such comparison may be undertaken in relation to a number of factors, such as, for example, employment levels or the number of enterprises. The map explorer tool allows users to produce maps according to questions that they would like to see answered. The detail is such that one can analyse not just according to a particular sector, but also according to a field of activity, and even a particular sphere of production.
- A download data function: The tax portal allows users to download the raw tax data which is stored as “.csv” files. It further provides download and printing functionality so that users can print out whole dashboards and maps with attached data that they may have customised to address issues and areas of interest.

The portal also provides documents explaining the data curation process and the methodologies used in collecting and organising the spatial data.¹⁷

Questions¹⁸

There is a need for a critical assessment of the head office effect, which can be quite significant – for example, in the banking sector in Johannesburg, in the retail sector in Cape Town, and generally in relation to the construction industry. This is important since the visualisations produced from the Spatial Tax Panel are being used in the presentation of integrated development plans (IDPs) by municipalities.

¹⁷ Following this presentation, participants at the meeting split into facilitated groups which explored the functionality of the online spatial tax data portal.

¹⁸ This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

The head office effect is also an issue with utilities. Efforts to quantify this more effectively continue.

Should race be a category in the Spatial Tax Panel given South Africa's history of spatial economic exclusion on the basis of race? Against this background, colour-blind data may be seen as reactionary.

Unfortunately, SARS does not provide data categorised by race, although it may be possible to infer this information using fuzzy logic. However, this represents a weakness in the administrative data – as does the absence of information on individual occupations.

9. REFLECTIONS¹⁹

The discussions that have been held have reinforced an understanding of the importance of administrative data used for statistical purposes and research, and in policy- and decision-making; and the importance of developing and disseminating administrative datasets further. In this regard, there are many lessons to be learned from the work undertaken by ONS in the UK. In addition, participants at the meeting have been introduced to the ongoing efforts undertaken by the Secure Data Facility and SEAD-SA to make such data available in South Africa. The past and future of the journey to promote the collection and use of such data, including the history of SA-TIED and the drive to produce an integrated data lake, have been presented, with the emphasis placed on the importance on collaboration – a community of practice – in advancing this work.

The discussions have also highlighted the importance of ensuring that the administrative data being produce is accessible and usable, which entails ensuring that the data is presented in a way that is accessible to users (for example, via the spatialdata.org.za portal); and providing training so that users are able to extract and use the data in meaningful ways, promoting their own understanding and that of others, and informing development efforts.

¹⁹ This section is based on comments made by Dr Ayanda Hlatshwayo, Chief Data Analytics Officer, National Treasury, at the end of the first day of the workshop.

10. LEARNING FROM EACH OTHER: UK CLUSTERING ANALYSIS OF SUBNATIONAL INDICATORS²⁰

ONS has clustered a number of key datasets relating to local economies, demographics, service delivery and health/wellbeing in an effort to:

- Indicate which local authorities are similar to each other in relation to their outcomes and demographics, so that local authorities have the information required to collaborate on shared challenges more effectively and learn from each other;
- Enable analysis of the trends that have led to local authorities being geographically grouped in particular clusters;
- Identify similar groups of local authorities so that more nuanced comparison/control groups are available for analysis, which can promote more effective evaluation of the impact of new devolution policies and provide improved evidence in support of place-based interventions; and
- Develop a suite of UK-wide data which can be used in ONS subnational products.

Following consultation at the local level across the UK about the indicators that would be used as part of the data clustering, a number of UK-wide datasets as well as data compiled from devolved administrations was curated for more than 300 local and unitary authorities across the country. With the aim of including all data considered useful for holistically understanding similarities between and among local authorities, data relating to 35 variables was collated.

This data was then processed deploying k-means clustering machine-learning so that the local authorities across the country were grouped together in a limited number of clusters indicating similar characteristics and performance in relation to the 34 variables. So, for example, certain local authorities may be grouped together in a cluster indicating higher connectivity (relatively strong delivery of services) and relatively low levels of health and wellbeing.

Having created the clusters, analysis by a number of key local characteristics can produce useful understanding on possible factors that may underpin the clustering. To this end, analysis of the clustering has been undertaken across England in relation to: urban/rural location; indices of multiple deprivation; region; age; population density; and whether the place is on the coast.

So, for example, 90% of local authorities where there is higher connectivity and lower levels of health and wellbeing are in urban areas (although urban areas only account for 57% of local authorities), while 60% of them are in the most deprived quintile (although only 21% of local authorities fall under this category). In other words, urban poverty may be viewed as driver of lower health and wellbeing notwithstanding relatively strong service delivery. Such analysis can help to clarify differential policy outcomes across clusters.

The clustering divides all the datasets compiled from the 34 variables that are used for assessing the performance and nature of the local authority areas into four models. In addition, there is a global model which includes all these indicators as well as a further one relating to the qualifications of working-age people. The four specific models, which address the economy; demography; health and wellbeing; and connectivity of the areas studied, focus on a number of key variables:

- The economic model includes measures of productivity and income, such as gross value-added (GVA) per hour worked; gross disposable household income; and numbers of children in relative poverty, as well as

²⁰ This section is based on presentations made by Emma Hickman, Deputy Director, Subnational Statistics and Analysis Division, ONS; and Jim Hawkins, Subnational Development Analyst, ONS.

employment and business-based information, such as the percentages of people employed in the construction, manufacturing and service sectors; and the number of active and new businesses. It also contains information on median house prices, which can indicate the wealth of a particular area.

- The demographic model includes census data on proportions of residents by ethnicity, age and religion, and on the density of and changes in population.
- The health and wellbeing model, which includes information on life expectancy by sex, and numbers of smokers, as well as data from surveys on levels of happiness, life satisfaction and anxiety.
- The connectivity model includes metrics relating to modes of transport (train, buses, walking and cycling); electricity consumption; CO2 emissions; and broadband capabilities and internet usage.

The data that is produced by clustering in relation to the four models can be presented for interpretation in a number of ways. For example, under the demographic model, clustering can produce the following four categories:

- Cluster 1: Older residents; predominately white; lower amounts of working age population (16-64); and people not of white ethnicity.
- Cluster 2: Younger residents; and high population change between 2011 and 2021.
- Cluster 3: High non-religious population; lower population change; and fairly average for other metrics.
- Cluster 4: High working age; younger population; high proportions of ethnic minorities, with lower proportions of people of white ethnicity; high population change between 2011 and 2021; and high religious population.

These categories may be analysed by local authority in the form of radar plot, or geographically, on a map, which indicates where in the UK each of these clustered kinds of demography will be found. In this example, “cluster 1” local authorities are generally found in English rural areas; “cluster 2” authorities are centred in Northern Ireland, southern England and a patch in north-west England; “cluster 3” authorities are centred around Scotland and south Wales; and “cluster 4” ones are in London and a couple of other English urban areas.

A key concern for ONS in developing the clustering tool was to ensure that the relatively complex analysis facilitated by the tool could be presented in a relevant and understandable way for non-technical users. To this end, it developed an interactive visualisation tool on the ONS website which allows users to access the data by local authority. The visualisation provides some basic information about the cluster within which the local authority has been grouped, including the demographic and geographic characteristics of the cluster. (ONS also provides the methodology, raw data and codes used for the clustering so that analysts with technical skills can recreate and adapt the analysis as they see fit.) ONS made an effort to name the clusters in ways that indicate what they represent without damaging the reputation of the local authorities that find themselves clustered under these headings (and thus causing offence that may damage relations between these authorities and ONS). In this regard, the headings for the clustering indicated whether the particular authorities were above or below the median for the four indicator models.

The clustering analysis has also been integrated into the Explore Subnational Statistics online offering to allow users to compare local authorities across indicators in relation to other similar authorities, as well as in relation to all authorities. So, for example, a local authority may compare local performance in relation to indicators such as gross disposable household income, gross median weekly pay and GVA per hour against that of all local authorities as well as against that of demographically similar ones. The information derived from such comparisons may inform the local authority’s efforts, particularly in relation to collaboration with other similar local authorities which may be overperforming or underperforming in relation to these metrics.

Such analysis has contributed to efforts to establish networks among local authorities which have been driven by the local authorities themselves with the support of ONS. The understanding of which local areas are similar to others enables local authorities to:

- Differentiate themselves from neighbouring local authority in evidence-based ways;
- Collaborate with others in the same cluster as themselves to address shared challenges;
- Learn how those in different clusters have improved their performance relative to other authorities in their cluster; and
- Identify relevant other local authorities for comparison when developing their own statistics.

From a national government perspective, the clustering outputs are particularly useful in providing evidence in support of effective locally targeted policymaking and spending and in evaluating the impacts of such efforts and whether they may be enacted more widely. In addition, the suite of local-level data that has been curated for the clustering initiative can be utilised by any analyst seeking to create their own data products and models. In this regard, ONS itself is also looking to use the clustering analysis in other areas of its work, for example, in relation to understanding access to amenities.

Questions²¹

Prior to the question-and-answer session, there was some discussion from the floor about data activities in South Africa that may be of value, such as:

- Standardising local surveys across the country in an effort to harmonise data production and cut the costs of this form of data collection;
- Leveraging data related to the delivery of services to residences; and
- Extracting data from travel surveys on places of residence and work, and, by inference, the provision of other services.

Would there be value in using more open data to complement the clustering data?

ONS has considered using satellite image information, although the relevance of such data in the UK, where the economy is largely service-based is not as great as it may be in South Africa, where the nature of economic activity is more diffuse.

Has ONS encountered criticism in relation to the images of places that were produced as a result of the ways in which the data was aggregated to make clusters? For example, London is portrayed as quite young and dynamic compared with other parts of the country, which may be viewed as an unhelpful stereotype.

There is a risk of such criticism; and, in this regard, ONS acknowledges the importance of granularity – and the ways in which there can be great disparities among regions and within local authorities which defy aggregation. At the same time, clustered data allows for quite nuanced comparison and facilitates effective evaluation of policy and economic efforts. In general, consultation about the variables being deployed and open communication can promote understanding of the benefits and limits of the analysis that can be fostered through clustering.

How were the clustering variables chosen and how was the problem of the use of different methodologies to produce the various datasets addressed?

²¹ This sub-section is based on a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

The original metrics for the clustering were derived from those used by the national government in its levelling-up programme. The indicators tended to measure outputs, such as productivity, household income and life expectancy. At the same time, there was an acknowledgement that the indicators would need to be compared in relation to a range of locational and demographic classifications if they were to be useful. In order to ensure the relevance of the variables, there was consultation with the devolved administrations across the UK and a process of correlating and winnowing variables was undertaken.

Customising the production of data according to local needs analyses offers an effective way of producing locally relevant information – but is there not also a role for harmonisation of local data across areas?

Local authorities tend to have a strong understanding of local geographies and the needs of local populations, which offers a starting point for conversations with other authorities which are apparently in a similar situation. Such conversations have value in terms of learning from one another, even if it found that the points of similarity are not as many or significant as had been thought.

There is a culture of working in silos in South Africa which can inhibit collaboration. How may this be addressed?

Collaboration is crucial when working with limited resources, so that efforts are not duplicated. The more that statisticians can push to break down silos and promote collaboration across South Africa, the more powerful the statistics, data and analysis being produced will become.

11. ENABLING EVIDENCE-BASED DECISION MAKING: OPPORTUNITIES FOR AND CHALLENGES TO BUILDING INTEGRATED DATA FLOWS

11.1 City of eThekweni²²

The Economic Development Unit in eThekweni Municipality has established a Special Economic Database, which is being used to connect multiple data sets with spatial indicators so that more comprehensive economic information can be made available for specific areas. The database comprises:

- Valuation roll data;
- Data on the consumption of electricity and water which requires quite a lot of cleaning;
- Data on Treasury-funded capital projects which features GPS (global positioning system) coordinates and the value of the investment being made;
- Data on spatial transformation trends derived from spatial development frameworks which includes information on applications for rezoning and building permits;
- Trade data extrapolated from certificates of origin (through a pilot project undertaken with the local chamber of commerce) which provides a picture of where imports come from and where exports are going, as well as the route taken – this data is useful as part of efforts to support African exporters in the city;
- Data from business and informal trade licensing, although the data on the informal sector is quite sparse ;
- Private sector data on opening hours and locations of 30,000 businesses across the city which offers a sense of the potential for new forms of activity, such as a night economy;
- GIS data on investment corridors and priorities;
- Earth Observation data indicating ward-level economic activity; and
- Data from a range of external sources including Stats SA, S&P Global Insight, Quantec, fDi Markets, a construction projects database and SEAD-SA, which provides a lot of time-series information at municipal, as well as sub-municipal, levels.

In collaboration with Innovate Durban, which is a non-profit company; local universities and the SDF at National Treasury, the municipality is seeking to integrate and disaggregate the data to hand at a centre called EDGE.GovLab. For example, the centre is digitising and integrating datasets on building applications, rezonings and economic incentives. In addition, it is seeking to facilitate research in relation to spatial economic data. The centre forms part of a Durban Edge project which offers an online portal through which users can explore the spatialised data produced, collated and held by the local authority. The portal offers a public-facing platform so that the data held by the municipality may be shared as widely as possible, while protecting sensitive information. The portal also offers a platform for internal use which provide data in a relatively raw form as well as in the form of dashboards on areas of interest. Data and code underpinning the offerings is available to be shared – and the hope is that other municipalities also will share such information and code, including cadastral GIS data.

The production of spatial economic data by the municipality and the SEAD-SA Spatial Tax Panel has facilitated useful comparisons with the other main cities in South Africa. For example, analysis of the data on employment by sector provides a clear idea of the relative strengths and weaknesses of Cape Town, Johannesburg and Durban

²² This sub-section is based on a presentation made by Justice Matarutse, Programme Manager for Innovation, City of eThekweni.

by sector, as well as their levels of economic activity and productivity more generally. The analysis shows that Cape Town has created relatively significant levels of employment in the retail and wholesale sector; while Johannesburg and eThekweni perform relatively well in the financial and professional services sector.

Spatial economic data is also useful for analysing a number of the fundamentals, such as income distribution and levels of business activity, underpinning municipal performance. For example, the spatial tax data reveals that the majority of the population in eThekweni are earning incomes below the SARS tax threshold (with 50% of those in full-time employment earning under R6,400 a month), while the numbers of those receiving public social grants are quite high. This income-level profiles may be seen as a disincentive for business – an assumption that is confirmed by data on the number of businesses compared with size of population, which ranks eThekweni as sixth among Metros, even behind Nelson Mandela Bay. The message for policymakers from this data is that there is a need to establish more firms and support the development of the informal sector in order to meet the demand for jobs among a growing urban population.

EThekweni also deploys a range of spatial economic data derived from administrative sources to promote understanding of economic activity and development at the local level within the city. Data from the Spatial Tax Panel is used to measure the impact of large-scale public sector capital projects in the places where they are being undertaken. Meanwhile, property-by-property data from the valuation roll is used to provide snapshots of residential growth and levels of business activity. When correlated with data on where businesses are opening and closing provided through the Spatial Tax Panel, this information can be used by municipal planners to identify where they should be focussing their research in support of efforts to boost economic activity and employment. The data, which can track levels of economic activity against the property rates being contributed to the public purse, can also be used to identify the wards where rates are not being collected properly.

Data from the valuation roll which tracks the number of new properties being established in an area, as well as the overall value of those in the business and commercial category, can be used to provide an evidence-based view of spatial growth and development trends – for example, in relation to the establishment of Umhlanga as a new central business district, and the failure of Amanzothi to become another CBD as had been anticipated. Data from the valuation roll is also being correlated with data purchased from private-sector providers on business and industrial activity to provide a picture of the South Durban basin, which is an area that has been decaying for some time and which is a target for economic rejuvenation.

Similarly, valuation roll data is analysed according to an algorithm checking property values, uses and sizes over time, as a basis for identifying properties and areas which are deteriorating. Building inspectors are then dispatched to these properties and areas to identify and visualise what is taking place – and this data is then used to inform potential policy interventions.

Other data used by the municipality to track economic activity, includes Earth Observation Data which indicates that significantly lower levels of economic activity in townships than elsewhere, despite the relatively dense populations in these areas.

Recognising the importance of integrating administrative data sets to produce useful knowledge, the municipality has sought to integrate the administrative data that it constantly produces as a result of its own transactions with other datasets with the aim of creating and visualising models that may be deployed by local officials. So, for example, there is a dashboard that project managers may use to select and filter a range of data for a particular area, including on: employment by field; the kinds of property in the area (residential, retail, commercial or industrial); and the scale and location of businesses in the area.

The data which provides a picture of land uses and the kind and extent of economic activity in an area over time can also be analysed to identify the factors that have led to a decrease or increase in GDP in that place – with the findings being used to inform zoning policy under the spatial development framework. In this regard, the use of

such spatial economic data has indicated how important the transport and storage sector is to economic development in eThekweni, as well as the relative importance of a range of other facilities and services to the urban economy.

The municipality also make uses of administrative data to inform decision- and policy-making in spheres not directly related to economic development. For example, road-crash data across time is used to identify hot spots where there is a need to improve traffic signalling or deploy metro traffic police. Meanwhile, social sentiment data pulled from Twitter is analysed and presented in the form of dashboards to the city manager and the mayor so that they have a better sense of the popular terrain in which they are making key policy and administrative decisions.

11.2 City of Cape Town²³

The municipal authority in Cape Town has placed great emphasis on the importance of evidence-based decision-making and has produced a relatively strong analysis of the city's economic performance over a number of years, leveraging official statistics (including mid-year population figures), as well as data provided by private providers such as Quantec and Global Insight, including estimates of city-wide GDP.

The data offers tracking of overall economic growth rates in Cape Town compared with those across the country. So, for example, it tracks the impact of shocks such as the Covid-19 pandemic, as well as a multi-year drought in the Cape Town area which appeared to cause less harm than had been expected (although there was some delayed impact). However, the data underpinning this tracking is derived from national data that has been heavily modelled and so may not be producing a true picture of the ways in which municipal and national growth trends differ.

Accordingly, the municipality deploys an array of high frequency administrative data sets that can be used to generate proxies for economic activity and can help to produce a clearer understanding of how the local economy is performing. In this regard, the municipality accesses:

- Data on passenger vehicle sales across the province, which indicates levels of consumer activity;
- Exports data from private providers, SARS and customs;
- Property development data in the form of building plans that have been passed (although this data may be more properly seen as an indicator of bureaucratic efficiency);
- Data on property prices, which indicates growth but also inflationary tendencies; and
- Data on electricity usage, which can act as a proxy for levels of manufacturing activity.

The municipality also leverages labour force survey findings produced by Stats SA to produce an understanding of employment and unemployment levels over time in the city, including in relation to national levels and the levels in other metros. This data indicates sustained growth in employment in Cape Town, particularly between 2017 and 2020. At the same time, there are questions over the statistical significance of QLFS findings at a metro level.

In this respect, the data provided by SARS, which may be considered more robust at the local level, has been of great use. For example, it confirms the QLFS finding that there was significant growth in employment from 2017 and 2020 and also shows how the city leads the field among metros in job creation in the formal sector. In addition, the SARS data provides more detail on the composition of the changes in employment levels by economic sector, including over the period of the Covid-19 pandemic. As a result of this tracking, it has been found that while employment in the retail sector has recovered well since the pandemic, employment in the manufacturing sector has struggled to recover.

The data also provides insights into what kinds of jobs in terms of wage bands are being created and where. So, for example, the data indicates that most of the growth in relation to full-time employment has taken place at the lower end of the wage spectrum, which is beneficial in terms of poverty alleviation. The data has further enabled analysis of job creation at the sub-metro level, indicating, for example, that while most of the new jobs that were created from 2014 to 2022 were situated in the CBD and along the N1 corridor, the greatest relative rate of growth in jobs created was in more outlying areas such as Kuils River, the Deep South and Helderberg.

At the same time, the data produced by SARS has its limitations, including, in particular, the absence of information on the informal sector and unemployment. In response, the municipality in Cape Town merges data sets, including

²³ This sub-section is based on a presentation made by Paul Court, Manager: Economic Analysis, Policy and Strategy Department, City of Cape Town.

those from SARS and those from the QLFS, when trying to understand broader labour market trends and producing its comparative labour statistics which detail unemployment and labour absorption rates.

The municipality in Cape Town has found the tax data particularly useful for in-depth analysis of economic activity in particular areas. For example, it has collated this data to provide a detailed picture of economic activity in the Atlantis special economic zone (SEZ) which it has shared with governments and firms planning to invest or investing in the area. The picture offers information on the percentages of floor space in Atlantis that are being used for manufacturing or other purposes, as well as the percentage that is vacant. It provides data on employment trends (which have largely recovered since the pandemic), including in relation how many people are being employed by sector. In addition, the employment data is available in relation to the age and gender of employees, which has shown that the employment of young men has not recovered since the pandemic (a finding that may be explained by the relatively sluggish recovery in the economic sectors in which this population group tends to work).

The kind of administrative data provided by the Spatial Tax Panel and other sources is not only useful for benchmarking statistics, and enabling analysis of economic performance in particular areas, as well as for comparisons of economic performance between and among areas – the data may also be deployed for applied analysis and research purposes as is illustrated by the following cases.

Case 1: Investigating copper theft

In an effort to establish determinants that may be exacerbating thefts of copper from electricity cables and installations, and in particular whether a recent ban on copper exports had made a significant difference in relation to the level of theft, the municipality accessed and correlated data on: repairs due to thefts provided by municipal asset managers; loadshedding times (it is easier to cut power cables and access the copper when the electricity supply has been turned off); seasonal effects; containers being shipped (as a proxy for economic activity); and, using SEAD-SA data, the proximity of scrap yards (under the assumption that thefts may be more likely near places where the stolen metal can be sold).

Case 2: Efficiency of service providers

The municipality undertook research to compare services provided by the local authority against those supplied by out-sourced providers in terms of their relative efficiency and cost effectiveness. Deploying the online service requests made by consumers logging details of inadequate delivery, which offers granular data, the municipality was able to identify complaints hot spots and then link these to the areas in which local authority and external service providers were operating.

Case 3: Mapping impacts of infrastructure investments and disruptions

The municipality has also used administrative data to develop models that consider the impacts of major decision-making or shocks. For example, in order to evaluate the impact of infrastructure investments (or major disruption), particularly in transport, the municipality would need to understand the interrelationship between the labour and property markets and how this adjusts in space in response to change. To this end, the municipality has developed a general equilibrium model that can track, say, the impacts of the establishment of a new transport infrastructure: from how this may reduce travel times; to how this would then make some places increasingly attractive in terms of employment and housing; which would lead to the movement of people to different places for work and accommodation; which, in turn, would place pressure on the supply-and-demand dynamics in the labour and housing markets (for example, leading to housing shortages in particular areas); which may then lead to rising property prices.

The data required to produce such an analytical model includes lots of information on travel patterns, as well as detailed data to understand employment and residential choices and the impacts of these in terms of property

supply and prices, as well as wages. Crucial datasets include those on population densities, which can be derived from census findings, and employment by place of work, which had previously been inferred from cellphone data (although telecommunications firms which provided this free under Covid-19 are now charging for it) but can now be provided by SARS data on employment densities.

Although macroeconomic data has many flaws, such as a little sensitivity to local factors, it can be of great use in providing markers of GDP or GVA. In this context, administrative data, which has much greater local sensitivity, can offer powerful proxy indicators of economic activity and complement the findings produced by macroeconomic estimates.

In general, local municipalities and state-owned enterprise (SOE) sources, such as Transnet, hold a wealth of relatively accessible administrative data that can be used to:

- Foster understanding of economic performance, including through benchmarking across areas;
- Support research that can offer in-depth insights into causal factors; and
- Construct models for considering the potential impacts of interventions, thus informing major decision- and policy-making.

In this respect, spatial data is of great importance, particularly given the highly spatially diverse character of South African cities. There can be no understanding of a city's integrated web of markets and residents without sufficiently disaggregated data; and multiple data sources enable triangulation which can underpin a richer analysis of trends.

Questions²⁴

What is the organisational readiness in terms of human resources and technological capacity to ensure the availability of appropriate data and the sustainability of data analysis? There is an issue of resources and capacity in terms of leveraging administrative data effectively at many municipalities.

Experience of working in municipal structures, where work-skills plans may be weak and a relatively large number of employees seem to have jobs for life, offers evidence of a relative lack of bureaucratic capacity. At the same time, there are pockets of great capacity for the production and deployment of administrative data, including at GIS and economic development units and in the departments responsible for water and electricity services, although collaboration among these pockets is quite low.

In this context, the establishment of a council-approved data strategy can indicate that there is political will to promote the importance of data- and evidence-based decision-making across the institution, which can translate into equipping more staff with the skills and understanding required to appreciate the value of administrative data, and the need to deploy and share it more effectively. There is also an argument for increasing the size of the quite small units working on data production and sharing given how effective these have already been in developing and promoting the use of administrative data.

In addition, municipalities should seek to foster an open-data ethos, encouraging practitioners in the field to share their data sources and applications with each other so that there is greater capacity across the system as a whole.

What data are the municipalities in eThekweni and Cape Town producing on climate change and renewable energy?

²⁴ This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

The eThekweni metro deploys GIS data to project climate-change; and also collects data on the industrial area in the South Durban Basin where there is a perennial risk of flooding, which makes insurers unwilling to provide cover for the businesses that are based there. Cape Town has projected the potential impacts of major flooding as part of a project modelling the impacts of extreme events.

In relation to renewable energy, eThekweni has conducted surveys on including independent power producers in the electricity grid; and its GIS unit has produced estimates of the amount of solar power that may be generated by property owners in relation to the size of their roofs. Meanwhile, Cape Town has undertaken a cost-benefit analysis of the returns on investment in solar photovoltaic (PV) plants.

What were the conclusions of Cape Town's study on the factors influencing levels of copper theft?

It was found that there was a correlation between levels of theft, and copper prices and the incidence of loadshedding, but no significant correlation in terms of the proximity of scrapyards. It was therefore noted that there was a need for greater vigilance and enforcement around the hotspots for copper theft when copper prices were high.

12. THE UK'S USE OF FLEXIBLE GEOGRAPHIES FOR ANALYSIS²⁵

One of the main aims of the subnational data strategy undertaken by the government statistical service in the UK is to produce more timely, granular and harmonised subnational statistics. To this end, a “flexible geography” approach may be deployed to allow users to build their own geographies at local level so that they can capture granular data in ways that make sense in terms of analysis and comparison. In the UK, the building blocks for such geographies comprise lower-level super output areas (LSOAs) or equivalent units. Each LSOA in England and Wales contains between 400 and 1,200 households. The LSOAs may be considered to be the UK equivalent of the local hexagons deployed by SEAD-SA in its spatial tax data panel. By grouping LSOA units together, users can create geographies beyond those offered by standard administrative geographies which can help them to analyse the spatial impacts of particular phenomena. However, it should be noted that the more granular the data, the more volatile it is likely to be; and, in this respect, the aim of the approach is to enable the establishment of geographies from which analytical conclusions may be safely drawn rather than, for example, to facilitate detailed comparison of one LSOA with another.

In the UK, a number of case studies have been undertaken using the flexible geographies approach to track how levels of gross value added may change locally in response to new conditions, including the construction of a new rail link; the onset of Covid-19; and the abolition of tolls on a major road bridge. GVA is a workplace-based measure of outputs against time that may be used as a proxy for economic productivity more broadly. However, given that it derives from workplaces rather than homes, levels of GVA can vary widely between LSOAs where there is significant industrial/retail activity and LSOAs which are primarily residential – indicating the importance of combining LSOAs to build aggregate areas for analysis rather than seeking to compare individual granular geographies directly and draw inferences on this basis. In 2023, ONS disaggregated GVA data by LSOA across England and Wales and has made this data available to use via its website and an API, allowing users to build their own flexible geographies using this indicator and census data. GVA data has been used as a basis for a number of case studies undertaken by ONS.

Case study: West Midlands Metro

The West Midlands Metro comprises a light railway system operating between Wolverhampton and Birmingham, which are two major urban centres in an area that was formerly England's industrial heartland. The rail line was built in 1999 and further extended in 2015 and 2019, enabling a time-series analysis of the changes in local gross value added as result of the extensions. In order to undertake this analysis, a flexible geography for the area through which the rail line passed was established by grouping together LSOAs around the line. Data from this area on GVA and other indicators including population numbers and house prices was then compared against similar data for the West Midlands region as a whole and the UK. It was found that from 2012, when it was announced that construction was beginning on extending the line, population growth in the immediate area outpaced that across the region and nationally, while local house price rises also outpaced those in the region and nationally. Meanwhile, after an initial dip, a surge in GVA growth brought the area around the Metro into close alignment with productivity levels across the region and nationally.

Case study: Impact of Covid-19 on subregional productivity

²⁵ This section is based on presentations made by Emma Hickman, Deputy Director, Subnational Statistics and Analysis Division, ONS; and Jim Hawkins, Subnational Development Analyst, ONS.

The West Midlands Metro case study indicates how data can be aggregated according to a particular geography to provide insights. Another approach is to drill down through aggregated data to understand how the data may be the product of local-level drivers and impact. This mode of analysis was deployed to identify and analyse the local economic realities driving falls in GVA during the Covid-19 pandemic. Reviewing the GVA data for 2020, it was found that the West Midlands had experienced the largest relative reduction in GVA across the UK. Drilling down into the West Midlands data, it was found that the Stratford-on-Avon had the largest fall in GVA. Analysis of the data on economic productivity in this local authority indicated a relatively high concentration of businesses in the science, research, engineering and technology sector, and a relatively high concentration of business head offices. Drilling down further, a particular zone, Lighthorne Heath, was identified as both the most productive area, as well as one hardest hit in terms of falling productivity. Analysis found a heavy concentration of automotive firms, including the head offices of high-end car manufacturers Aston Martin and Jaguar, in this area – indicating the local factors that may be driving the drop in productivity at this time (although given the unreliability of granular data, any such inferred conclusion would need to be bolstered by on-the-ground research).

Such top-down use of flexible geographies can be useful as a way of:

- Starting to explore a topic (in this case the impacts of the pandemic on GVA);
- Generating hypotheses for analysing statistics more generally (in this case, the hypothesis that small economies tied to a few large firms are more vulnerable to shocks);
- Offering a model for testing the hypothesis (in this case, whether there is correlation between relatively homogenous economies and a lack of resilience in terms of productivity); and
- Establishing the direction for building bespoke geographies for further comparative analysis.

Case study: Severn Bridge tolls

This case study concerns an evaluation of the impact of the removal of tolls on the Severn Bridge, which is a major road link connecting the port of Bristol and south-west England with south Wales, on the GVA of businesses in south Wales. A survey conducted some years before the removal of the tolls in 2018 had indicated that 50% of firms in south Wales considered the impact of the tolls to be an important or very important factor for their businesses.

In order to test the actual importance of the tolls in the light of their removal, a study was established deploying a “synthetic control” methodology. Under this methodology, a “treatment area” is identified – in this case, local areas with high concentrations of employment in Newport and Monmouthshire, situated near the Welsh end of the bridge. Then a number of “control” local areas are selected, which should be similar to the “treated” ones – in this case, the areas had to be at least 90km away from the bridge and were chosen on the basis of their concentrations of employment as well other characteristics including the presence of business parks (which was a feature of the “treated” areas).

The analysis then compared the change in GVA in the areas in south Wales against those in the control group and found that there was no significant difference between the trajectories. In response to this finding, the analysts then considered other factors, such as rising house prices; and it was found that, even before the bridge tolls were removed in 2018, there was a clear divergence between what was happening in the control group versus what was happening in Newport, where house prices increased quite a lot.

The main economist in the team undertaking this study noted that this was an important finding – indicating that the gains of the policy intervention (removing the tolls) had been channelled into the capital base of home owners rather than into increased productivity. Given that this had not been the intention of the policy, it was recommended that further such interventions should include building additional housing stock to ensure more democratic distribution of the economic benefits that may be accrued.

Questions²⁶

There is clear value in adopting the flexible geographies approach in South Africa, deploying the hexagons in the Spatial Tax Panel data while bearing in mind the unreliability of such data at the granular level. But how can the country produce an indicator for comparison and analysis at the local level, such as that provided by GVA?

The UK produces structurally sound surveys that provide GVA at a national level, which is then apportioned down to local authority level by regional teams who link the information to VAT data. In addition, there is data from the national business register that is available by “enterprise group” and also provides detail on turnover and employees at local units, which allows calculation of GVA by firm level and also extrapolation of this information at the level of LSOA.

In the analysis of flexible geographies, it seems that London is considered to be a unique case. Whis is this so and how does it affect national government policy-making?

London is the national hub for financial services, which is probably the most productive sector of the national economy. The city is also more ethnically diverse than other parts of the country and derives greater conglomeration benefits than anywhere else. So, its productivity is relatively high – and even though it is not performing as well as previously, it still outstrips expectations on a range of measures. In this context, government policy and investment has tended to focus on the secondary cities which are underperforming by comparison.

Under the terms of the 2024 United Nations Climate Change Conference (COP29), can air-quality levels be compared across cities and places in cities?

Yes, and it would be quite possible to build flexible geographies on the basis of GVA or income to compare CO2 emissions by area, including in relation to other factors such as the nature of the built environment and the presence of particular facilities in these places.

²⁶ This sub-section is based a plenary discussion, with comments and questions from the floor and answers offered by the presenters.

13. INTRODUCTION TO ONS INTERNATIONAL DATA MASTERCLASS²⁷

The masterclass is an online training resource comprising six hours of learning and offering a certificate that aims to support policymakers in placing data at the heart of public decision-making. The course has been developed by ONS, the UK Foreign Office, the UN Statistics Division and the UN's Economic and Social Commission for Asia and the Pacific (ESCAP) in response to major challenges that were identified around the understanding, interpretation, use and presentation of data under Covid-19.

The masterclass aims to support and promoted understanding of how to:

- Use data and evidence to improve decision making;
- Use data in policymaking;
- Deploy new data-science methods;
- Support a data culture; and
- Communicate data narratives including in the form of visualisations.

It comprises three modules presented by senior statisticians and academics from the UK, Africa, South America and Asia on:

- How data is used to make policy, with an international example relating to Rwandan economic policy;
- Communicating compelling narratives through data, with an example of visualisation deployed by the British Broadcasting Corporation (BBC); and
- Data science and new statistical methodologies such as reproducible analytical pipelines, with examples drawn from Vanuatu and Belize.

The masterclass aims to produce a number of key learning outcomes. It seeks to show:

- How data science can be deployed for resource-intensive tasks, supporting faster, better use of resources;
- How good quality data lies at the heart of robust decision-making;
- How data may be used to diagnose problems, and produce targeted interventions; and
- How data visualisations should be used and interpreted.

²⁷ This sub-section is based on a presentation made by Adil Deedat, Head of International Development Operations, UK:ONS.

14. PLANNING FOR THE COMMUNITY OF PRACTICE²⁸

What is most exciting about a community of practice on the economy of cities?

Those attending the workshop identified a number of benefits that may be derived from participation in the Community of Practice in terms of:

Knowledge to be gained

- Municipalities can learn lessons from the work undertaken in other cities;
- Understanding of the spatial dynamics of urban economies may be enhanced;
- New, valuable data sources for research and to inform policy- and decision-making in government and industry may be shared;
- Lessons can be learned from projects and innovations that have been developed elsewhere, for example, in the UK;
- Access to indicators can inform and assist in planning from ward to metro level;
- The knowledge and understanding derived from analysis of use cases can be of great practical value; and
- Understanding of the availability and value of granular data at the local level and how this can complement national data may be promoted.

Collaboration to promote effective data-sharing and use

- Knowledge, experiences and best practices in relation to data on spatial dynamics may be shared;
- A joint problem-solving approach, perhaps including crowdsourcing mechanisms, may be adopted so that solutions can be found collaboratively;
- The value of data-sharing and production will be enhanced through the engagement of multiple stakeholders, including across the public and private sectors, and among academics and policy-level stakeholders; and
- Collaboration may foster discussion on, and a broader appreciation of, the value of sharing spatial economic data.

Bureaucratic benefits

- Collaboration to break down silos would foster greater bureaucratic efficiency;
- Re-using work in the form of administrative data that has already been produced represents an efficient form of data use;
- Fostering bottom-up innovations would allow for development of a view of what present proficiencies are; and what user preferences in present workspace are, as opposed to having those things dictated by a more senior authority.

Improved practices

- Data practices at the local and city level may be integrated into those at the national level;
- Local data practices would be benchmarked against the best practices at the global level; and
- The professionalisation of government would be fostered by effective data-sharing in support of evidence-based decision-making.

²⁸ This section is based on feedback from four breakaway groups on this topic.

Educational resources

- External resources that may be leveraged to address internal problems will be identified;
- There will be access to a range of expertise for exploring data; and
- It will be possible to keep up-to-date with the latest technologies relating to data production and analysis.

Personal development

- There will be value in networking among members of the community; and
- Valuable lessons may be learned from the successes *and* failures of other working in the same field.

Which topics/themes are of greatest interest?

Those attending the workshop identified various topics of interest and a number of benefits, needs and concerns in relation to these, including:

Actual and potential benefits

- The value of sharing new technologies, methodologies and best practices;
- The value of the provision of high-frequency data;
- The value to users of a range of different data sets relating to economic activity, foreign direct investments and regional growth;
- The usefulness of flexible geographies and the need to produce more tangible data in support of such analysis; and
- The importance of being able to access information on latest technologies;

Needs/concerns

- The issue of how best to ensure the efficient production of accurate, valid data;
- Issues around public- and private-sector indicators for the value of property (rates and market prices), which can be out of kilter;
- The issue of how best to promote the prioritisation of data-analysis budgets in the public sector;
- A need to undertake spatialised economic modelling at sectoral and departmental scales as well as at the metro level;
- A need to develop standard units for spatial analysis for different sectors (education, healthcare, etc) to enable comparisons between different parts of the country;
- A need for environmental modelling to promote understanding of climate-change impacts;
- A need to address the strategic aspects of change management, for example, in relation to greater inter-departmental and intra-institutional collaboration, in support of more effective data production and sharing;
- The issue of what kinds of reporting lines may be established between local authorities and national government in relation to data production and sharing;
- The issue of how best to support municipalities that are early in their data maturity so that they can develop their capacities for using and analysing data;
- A need to ensure that no one is left behind, by engaging external data providers and participants in the informal economy so that they too can derive benefits from the work undertaken by the Community of Practice;

- A need to ensure that universities and technical colleges are engaged in the work of the Community of Practice so that their knowledge-production agendas are responsive to the problems faced by communities. (To this end, the research and data that is being produced should be co-created by the relevant stakeholders);
- A need to promote the beneficial impacts of using and sharing spatial economic data;
- The issue of how best to communicate data so that it can influence policymaking;
- A need to raise awareness, including among communities, about the work that's being done; and
- A need to develop further ways of sharing the work that is being undertaken so that the importance of spatial economic data may be promoted more widely.

What sorts of activities should be pursued by the Community of Practice?

Those attending the workshop identified a number of kinds of activities to be undertaken:

Meetings

- A conference should be convened to discuss and find ways of leveraging the impact of data analysis in local government;
- There should be continual meetings among members of the Community of Practice so that momentum is maintained and topics of interest are explored;
- Quarterly meetings may be held by different subgroups to discuss matters of interest and their work; and
- A range of sessions should be convened to discuss use cases and share experiences.

Training and research

- There is a need to re-skill and up-skill those working in the sector, including in terms of inculcating digital skills for research. Indigenous knowledge also should inform data practices.
- Training should be provided in data analysis and visualisation; the use of AI in data analytics, machine learning; and data-sharing best practices;
- Training engagements should be designed in a tiered way to meet the different needs of the individuals in the Community of Practice, whose levels of expertise may range from beginner to expert;
- Training should be offered online to promote access and save money;
- Training and workshop activities must speak to a diverse range of stakeholders and address problems faced by communities; and
- Research on spatial economic data, including work in progress, should be shared and peer-reviewed with the community so that it may be improved.

Promoting engagement

- Young people and undergraduates must be included in Community of Practice activities so that they are introduced to the ways in which data can be used to inform decisions. This may be undertaken through site visits, hackathons and data-thons;
- A short two-minute tutorial on how the Spatial Tax Panel and dashboards work should be produced so that practitioners can quickly learn how to use these tools; and
- Data projects and activities may be communicated to the Community of Practice via social media, such as Instagram.

15. CLOSING REMARKS²⁹

The SEAD-SA Community of Practice offers a compelling proposition which may be summarised by its slogan describing its aim to facilitate the use of “data for inclusive and vibrant city economies”. In the context of massive unemployment, job creation is a top priority in South Africa. Given that the metros comprise the bulk of the national economy, they have a particular responsibility to create more jobs by delivering on their potential for growth. In support of this agenda, efforts should be made to generate a greater understanding of what shapes the national and urban economies, including the factors that may be inhibiting growth or that may help to foster it – and to promote the importance of this mission. At present, the government seems to be focussed on Eskom and its problems; Transnet and the logistics crisis; and crime. By contrast, the Community of Practice is focussed on understanding how to turn the economy around so that there is more employment. Such an understanding depends on producing and sharing better knowledge on the drivers of, and impediments to, economic growth in cities, which may include their physical assets and infrastructural constraints; local levels of economic activity, investment and innovation; financial and technological provision and capacity; and the human capabilities that help to generate growth, including people’s levels of education, and their skills, knowledge and talent.

At present, there is insufficient understanding about the drivers and constraints in large part because of gaps in the data, and the fragmentation of the knowledge that has been produced, with different entities hold different bits of data. Meanwhile, however, there are great quantities of administrative data available that has been collected for all sorts of purposes – data which could be leveraged to foster greater understanding if it were accessible. Such data could be leveraged to provide timelier, cheaper and relatively easily produced information and statistics in support of efforts to act upon the priorities facing the country’s urban economies. In addition, a range of techniques deploying technologies such as artificial intelligence, machine learning and leading-edge software have created new opportunities for leveraging data in support of economically sound, evidence-based policy- and decision-making. In other words, there is a need to be more creative and energetic around the ways in which data may be produced, accessed and used so that the potential of this information for fostering economic development can be realised.

At the same time, it should be noted that the goal of realising the potential inherent in the data cannot be achieved overnight. It will take time to improve awareness of the importance of administrative data sources and to build trust among the different entities so that they share this data. In this respect, the establishment of the SEAD-SA tax data portal represents a good start; while the wide uptake of and interest in the administrative data provided by SARS indicates the benefits that can accrue from being prepared to share data. Those working in this field may also be reassured by the great steps taken by ONS in the UK, which now offers responsive, user friendly provision of administrative data at the local level, where once it provided little reliable economic data for cities.

In this regard, there is a potentially significant role to be played by Stats SA in championing a similar transformation in the provision of administrative data in South Africa – a role that it may fulfil by:

- Acting as a repository for the data;
- Providing quality assurance, thus promoting trust among users of the data;
- Helping to safeguard data against hacking and other threats, thus promoting trust among potential data providers who may otherwise be reluctant to share the information they hold; and
- Facilitating capacity building to support local authority and academic use of the data.

²⁹ This section is based on comments made by Prof Ivan Turok. Department of Science and Innovation/National Research Foundation (DSI/NRF) Chair in City-region Economies, UFS; and Distinguished Research Fellow, HSRC.

This central coordinating role should be complemented by active engagement on the part of the local authorities, government entities, researchers, and civil-society and private-sector stakeholders, such as consultants, seeking to access the data. In this regard, it is to be hoped that the Community of Practice can collaborate closely with all the relevant stakeholders at the central and local levels, learning about and sharing the data that is available, and improve access to it over time.

Given the scale of the mission to provide and share usable administrative data – and the challenges that may be faced along the way – it is necessary to adopt a staged agenda. The steps that should be taken to ensure progress and momentum must be planned. To this end, a number of questions need to be asked and answered, such as: What are the main data priorities? Should the emphasis be placed on providing new data sets, or on promoting new ways of organising the data, such as through flexible geographies and clustering similar areas, in order to meet municipal needs? Perhaps the approach should be a bit of both.

The project will also entail taking risks. Sharing data makes people nervous, but the benefits are great and the rewards make the effort worthwhile, as has been shown in the case of the data shared by SARS. In this regard, however, the initiative should also undertake to do more than just seek access to a wider range of datasets. It should seek to foster high standards and robust processes, procedures and systems to protect the integrity of the data being shared. In this regard, efforts to cut corners or inappropriate actions may jeopardise the initiative and stall its momentum.

That said, there is a need to be pragmatic rather than perfectionist about the quality of the administrative data being used, which has, after all, been developed for other purposes. In this context, the goals should be to develop standards and safeguards that are appropriate to the kind and form of the data being accessed, rather than unrealistic standards that can never be met.

In addition, it is important to consider and support the whole value chain – not only obtaining access to data, but also fostering uptake by:

- Making it accessible through mechanisms such as data portals;
- Building capabilities to access and use the data; and
- Encouraging links to researchers.

The goal must be to ensure that the data generates useful knowledge and provides strong evidence for better decision- and policy-making. In this respect, the Community of Practice is seeking to enact a virtuous circle, promoting better understanding of data users' priorities and needs, which shapes what data should be provided and how, which, in turn, ensures the relevance and quality of the statistics and information provided and their wide uptake.

The Community of Practice offers a space for mutual learning and also a site for promoting and pursuing a larger agenda, which is to support municipalities in developing the crucial role played by cities in the national economy so that jobs are created and the issue of unemployment is meaningfully addressed. Given that the Community of Practice is seeking to make a difference on a matter of such great national significance, it is well-placed to promote and advance its mission.