THE CHALLENGE OF SOCIO-ECONOMIC DATA FOR THE WESTERN CAPE TRAVEL AND LAND USE MODEL

F CONTIERO[1] and WG ALLEN, Jr., PE[2]

[1]Bentley Systems, Inc., Carl-Zeiss-Ring 5, 85737 Ismaning, Germany
Tel: +49 1728837391; Email: Filippo.Contiero@bentley.com
[2]Bentley Systems, Inc., PO Box 390, Windsor, SC 29856 USA
Tel: +1 803 270-7114; Email: Bill.Allen@bentley.com

ABSTRACT

Transport and land use models are famously data-intensive, requiring huge amounts of data on demographic and economic conditions at a high level of geographic detail. This information is needed for model development, calibration, and application. The new Western Cape Land Use and Transport Interaction (WCLUTI) model was especially ambitious, as it represents combined and connected models for transport and land use.

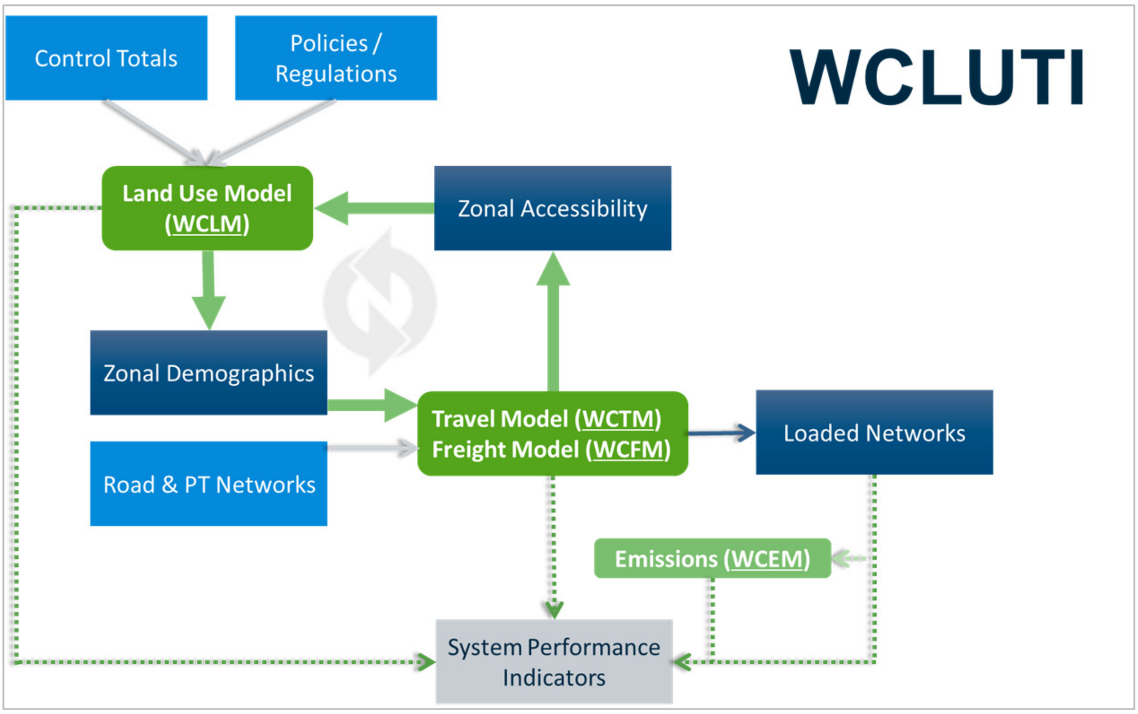
Transport models require information on the demographics of travellers, since most of their travel choices are influenced by their socioeconomic conditions. Land use models require even greater amounts of information, in order to accurately describe the supply, demand, and pricing of land and the relationships among those elements. Obtaining the necessary data at a high level of geographic detail is always challenging. Problems include public agencies that may be reluctant to share information, inconsistent data collection, and data that exists in different forms and for different years.

The Western Cape project addressed these issues by assembling the most accurate possible picture of observed travel and development from many different sources. Both public and private data sources were used. A large effort was spent on reducing inconsistencies between datasets. For example, satellite imagery and GIS tools were used to establish equivalencies between population and employment data at the Small Area Layer and at the Enumeration District layer. The data that was finally used represents an amalgam of different information, representing different geographies and different years.

In addition, the data being used for this model is never considered “final”. As new information is made available, earlier data is made more detailed and accurate. This is a continuous activity that will be built into the model development process.

1. BACKGROUND

Transport and land use models require huge amounts of information for model development, calibration, and application. A new integrated land use and transportation model, the Western Cape Land Use and Transport Interaction (WCLUTI) model has recently been developed, as has been documented in other SATC papers and the project report (1). This is a complete model of travel demand and land development, as illustrated in Figure 1. The creation of this model was especially ambitious, representing combined and connected models for transport and land use. During the first year of model development the initial version of the socio-economic database was created, based on the information available early in the project. Problems with these data were noted, and the validation of the goods movement model and traffic assignments, and preliminary scenario testing suggested that the socio-economic figures needed refinement. During the second year of model development additional information has become available. This paper explains the motivations and methodology undertaken to estimate more suitable socio-economic data for the WC Province, to be used in the WCLUTI model.



Source: Bentley Systems

Figure 1: WCLUTI Structure

The objective was to create a database of socio-economic information needed by the travel demand and land use models with spatial aggregation at their zonal level. It is important to highlight that the socio-economic data are necessary for the entire province at a finer level of aggregation than district or municipality because the WCLUTI model is a "multi-resolution" model, allowing for the modelling of detailed land use and transport both at the provincial and local scales, with spatial aggregation at:

- Small Area Layer (SAL) level – defined by Stats SA and extended for the model to include empty areas, used to summarise land use primarily for the Western Cape Land Use Model (WCLM).
- Transportation Analysis Zone (TAZ) level – created for the Western Cape Transport and Freight Models (WCTM, WCFM), for purposes of summarising trip-making.

Since a key function of WCLM is to provide input data to the travel models, the equivalency between these two layers was established.

Therefore, the following socio-economic data are necessary for the calibration, validation, and application of the WCLUTI model in the ***reference* Base Year 2016**:

- *Employment* data (jobs) at SAL level, stratified by economic activity.
- *Households and Populations* data at SAL level, stratified by the models' categories (income, size, dwelling type and size).

The main reasons why these data needed to be estimated from multiple available sources were:

- Unavailability of a unique official source of cleaned and verified data.
- Inconsistencies between the data and the actual land use within the data sources.
- Inconsistencies amongst the different data sources (e.g., totals not matching, different data definitions).
- Available data not reflecting any real census or counting of employment, but information pieced together from various estimates of employment with little information on the methodology used by the original sources.
- The different available sources represent data from different years.

The adopted methodology does not involve any statistical modelling of synthetic data, but it was rather oriented towards a "data assembly" approach, trying not to alter the original data sources, that were already mostly coming from previous estimations. The data that was finally used represents an amalgam of different information, representing different geographies and different years.

A large effort was spent on reducing inconsistencies between datasets. For example, satellite imagery and GIS tools were used to establish equivalencies between population and employment data at the SAL and Enumeration Area (EA) level, to address local issues.

The objective of this paper is not to provide a general framework for dealing with all these data inconsistencies. The purpose of this paper is to highlight the challenges faced in obtaining data suitable for the purpose of the model development, with focus on the needed actions for adjusting these data, trying to make use of the different data sources.

2. TERRITORIAL AGGREGATION LEVELS FOR WESTERN CAPE DATA

Data are provided in Western Cape according to different levels of aggregation, depending on the data source. The following are the number of zones for relevant levels of aggregation of the WC territory starting from the higher level of aggregation and going down to the finer level of disaggregation:

- Districts = 6
- Municipalities = 25
- Sub-Places (SP) = 1,578
- Mesozones = 2,937 (additional layer used by CSIR with estimated employment data)
- Small Area Layer (SAL) = 8,408
- Enumeration Areas (EA) = 10,818
- Land Parcels (LP) = 1,296,289

EA and SAL zonal resolutions are the most important levels because they are the levels of the Census data collection and distribution, they are the smallest levels that can be used for data aggregation and analysis and they are consistent in terms of boundaries between each other. Therefore, it was decided to adopt these two layers as reference zonal layers, and to focus on the SAL level for data aggregation because most of the data is available at that level.

Nevertheless, five major problems affect these zonal layers:

1. Boundary/spatial inconsistency amongst different levels (e.g., MesoZone layer and any other level, Land Parcel and any other level).
2. These zoning systems are not continuously covering the entire region, but “holes” are left for non-residential development areas at the time of the survey.
3. EA and SAL zones are often not aligned with the Land Use of the territory, its transportation system and its natural barriers.
4. EA and SAL zones spatial layout only considers the residential component of the Land Use.
5. The Land Parcel (LP) layer is not unique. Different sources provided different alignment and boundaries, making it complicated to create a unique correspondence between LP and EA/SAL, and to process General Valuation Rolls data (necessary for WCLM).

This issue of “holes” in the SAL zoning system (Figure 2) has been resolved by creating “dummy” zones within these empty areas. This results in a system of a total of 8798 SAL zones being defined for WC. It is recommended that a continuous zoning system should be used for Census data, with zones all across the territory (even where the data is null), for consistency in different years and in terms of having a common layer for different type of analysis.

From the Stats SA website, Enumeration Areas (EAs) are defined as “the smallest geographical unit (piece of land) into which the country is divided for enumeration purposes. Enumeration areas contain between 100 to 250 households”. Because these areas were defined for population enumeration purposes it is often possible to find cases in which the EAs/SALs do not neatly define or follow residential and employment clusters, natural barriers, or the road and rail systems.

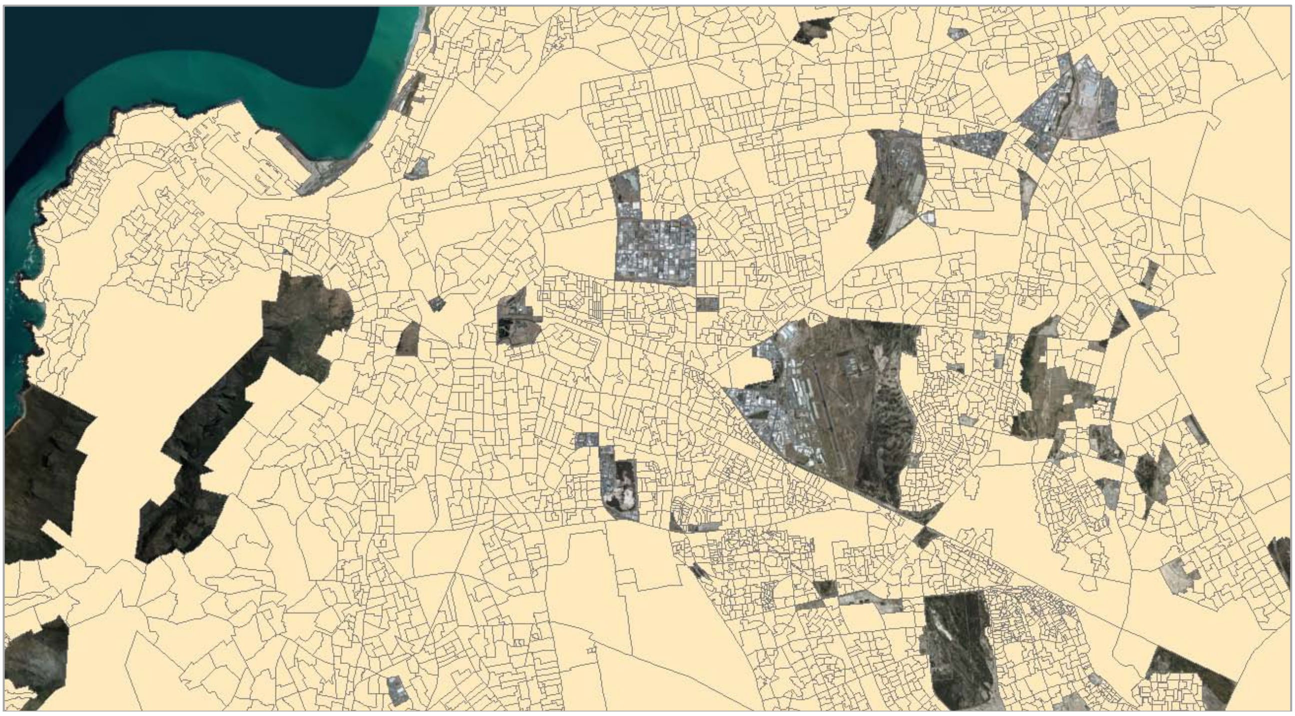


Figure 2: “Holes” in the StatsSA SAL zoning layers

3. EMPLOYMENT DATA

It is important to start the discussion by highlighting the key difference between:

- *Household-based data, or "origin-based"*, quantifying the labour force in an area in terms of employed residents.
- *Enterprise-based data, or "destination-based"*, quantifying the employment in an area in terms of actual jobs being undertaken in that area.

In other words, household-based data are not representative of the employment in terms of jobs in the zone because they are "origin-based", i.e., identifying where the worker *lives*, rather than "destination-based", i.e., identifying the number of employees where the employees *work*.

The three components of the WCLUTI model (WCLM, WCTM and WCFM) require "destination-based" employment (jobs) data by zone:

- WCLM: to identify the location choices of firms/jobs at SAL level in the base year, stratified by economic activity, for the model estimation and calibration, then, as control totals for the modelled regions (inside Cape Town, outside Cape Town), for the same economic activities, to run the model in "forecasting mode".
- WCTM and WCFM: the level of attractiveness of a zone is typically described largely in terms of employment. This refers to the number of full-time jobs in each zone (or full-time equivalents). This is of obvious importance in describing work travel, but it is also an important factor in determining travel for other purposes, such as shopping or personal business. So, it is important to know not just the total number of jobs, but the number of jobs stratified by job type. This relates primarily to the nature of the work itself, and not the employer or the employee. Note that the model classifies *jobs*, not *employers*. Having more categories of employment produces greater accuracy, as the trip rates per employee vary by job type, but it is vital to consider the trade-off of detail vs. user input when forecasting future scenarios.

3.1 Employment Data Sources

The main sources of employment related data are summarised below. The sources that are "greyed out" in the list are only reported for completeness but they werenot used to derive the employment data.

- Aggregate employment data at some level of spatial aggregation:
    - Household-Based: **SEP** (Socio Economic Profiles) derived from STATS SA QUANTEC/**QLFS** (Quarterly Labour Force Survey) aggregated at ***municipality*** level, by 10 economic categories: Agriculture, Mining, Manufacturing, Utilities, Construction, Trade, Transport, Finance, Institutional, Community (WC, 2015).
    - Enterprise-Based: STATS SA **QES** (Quarterly Employment Statistics) aggregated at **country** level (WC, 2016-2019).
    - Enterprise-Based: GeoTerra Image (**GTI)** data at ***EA*** level (D_MIGR_IN = estimation of total number of formal workers that "migrate" into an EA to work) without classification (WC, 2015).
    - Enterprise-Based: **CSIR** Geospatial Analysis Platform (GAP), estimated employment data at ***MesoZone*** level per SIC code (WC, 2009 and 2016).
    - Enterprise-Based: **MTIIF** estimation from **City of Cape Town** at CCT ***TAZ*** level, based on employment over Gross Leasable Area (GLA) ratio and other assumptions (CCT, 2015) – not used.

- Disaggregate employment data at the establishment level:
  - Braby's Business Database (WC, 2019) – sample size = 45,724 establishments; 228,597 employees – not used.
  - Employment Survey for the City of Cape Town (CCT, 2018) – sample size = 8,218 establishments; 226,747 employees – not used.
  - **Education**: Classrooms, Learners and Staff per School from WCG Education Department (WC, 2016), we summarised them at SAL level – total = 2,566 establishments; 53,545 total employees; 1,138,345 total learners.

- Auxiliary data on land-use/establishments dimension and location
  - Gross Leasable Areas (**GLA**) from General Valuation Rolls (**GVR**) defined at **Land Parcel** level with DOR and SAR codes (CCT, 2015).
  - **Agriculture**: disaggregate areas of agriculture and livestock, by sector, from WCG Department of Agriculture (2015-2018).
  - **Mining**: disaggregate mining extent areas from ESRI SA (1990, 2013, 2014).

From this list, it can be observed that no unique/official source of employment data with a useful level of zonal aggregation (SAL) could be obtained from the above sources directly.

3.1.1 SEP 2015 (Municipality – WC)

It has not been possible to use the QES data derived from enterprise-based employment surveys because of missing economic sectors causing the number of jobs to be much lower than QLS employees, comparing totals at South Africa level (QES is only available at national level).

The only suitable official data source in terms of employment at the municipality level and Standard Industrial Classification (SIC) for Western Cape was the Socio Economic Profiles. A necessary assumption was that these labour force figures (household-based) from SEP could be considered equivalent to employment data in terms of jobs (enterprise-based) to form our control totals at municipality level (to be further disaggregated at SAL level), ignoring inter-municipal commuting. Furthermore, data available from SEP 2017 gave 2015 employment values, nevertheless, these have been assumed to be sufficiently close to our reference 2016 values. Moreover, data quality uncertainty suggested not to apply adjustment factors, to avoid the risk of further increasing distortion in the data.

The definition of the SIC codes is provided in Table 1, together with the estimated number of jobs from WCG SEP for CCT and the entire WC.

Table 1: STATS SA SIC Codes with estimated total employment from SEP ("formal + informal" targets 2016)

SIC code	Sector	Description	Cape Town	Western Cape
1	Primary	Agriculture, Forestry, Fisheries	38,858	227,202
2	Primary	Mining, Quarrying	819	1,577
3	Secondary	Manufacturing	167,231	242,008
4	Secondary	Electricity, Gas, Water	5,770	8,309
5	Secondary	Construction	130,018	197,481
6	Tertiary	Wholesale/Retail Trade, Catering, Accommodation	382,476	590,015
7	Tertiary	Transportation, Storage, Communication	96,415	139,801
8	Tertiary	Finance, Insurance, Real Estate and Business Services	309,114	427,413
9	Tertiary	Community, Social and Personal Services, Public Administration including General Government	432,596	642,782
		Totals	***1,563,297***	***2,476,588***

3.1.2 GTI 2015 (EA – WC)

GeoTerra Image (GTI) provided a database containing demographic data at the EA level for 2015 (2). This database is a useful source of information, but it represents an *estimation* of demographic data derived by statistical analysis of different sources from different temporal scales. The database contains anomalies and some of the information should be considered more as comparative numbers, rather than absolute values. Referring to daily migration patterns, the underlying assumptions behind GTI information is that no inter-municipal migration is taken into account, in line with our assumption above for the target municipality totals.

The **D_MIGR_IN** values in the GTI database represent the estimated total number of *"formal + domestic workers"* that could migrate (i.e., travel) into an EA during a working day. Detailed information on how this attribute is estimated from GTI was not available in the source documentation. Several data inconsistencies were detected within the City of Cape Town, whilst more consistent information was generally provided outside CCT. Therefore D_MIGR_IN numbers have been factored by municipality (single factor by municipality and SIC code) to match the 2016 Stats SA targets and have been assumed to provide the total employment distribution for the WC province, used as main data source **outside CCT**. Within CCT, this attribute has been only used as reference for correcting data for some SAL zones. The D_MIGR_IN attribute does not provide information on the economic sector of the employment.

"Land use activities" were grouped by GTI into ten classes (AGRIC, FORESTRY, CONSERV, MINING, RESIDENT, COMMERCIAL, INDUSTRIAL, OTHER_BU, WATER, UNCLASSIFI) and reported in terms of hectares covered per EA. These attributes have been used to define the four land use proportions for each EA: agriculture, mining, industrial, other (NR_AGRI, NR_MINI, NR_INDU, NR_OTH).

The D_EDU field values in the dataset represent the *"estimated learners in pre-school, primary, secondary, and other schools"* in an EA during the day (learners at tertiary institutions not included). From comparison with the observed number of learners aggregated at EA level from the WCG Education Department data, large inconsistencies have been detected. Therefore, this second source has been used instead.

3.1.3 CSIR 2009 and 2016 (MesoZone – WC)

CSIR derived employment data at MesoZone level and SIC code from QUANTEC (consultancy firm providing economic and financial data in South Africa) household-based employment and Gross Value Added (GVA) data at municipality level. Table 2 gives the summary of 2009 and 2016 employment by SIC category from CSIR for WC. These overall figures do not appear to be the same statistics in terms of magnitude, and the 2016 data seems different by an order of magnitude compared to the WC totals from Table 2. Also looking at the proportions by municipality or at the individual MesoZone values for 2009 and 2016, these values seem very different.

Table 2: CSIR Employment by SIC code for WC in 2009 and 2016

Year	SIC_1	SIC_2	SIC_3	SIC_4	SIC_5	SIC_6	SIC_7	SIC_8	SIC_9	***TOTAL***
2009	108,483	4166.226	209,522	6124.014	No est.	288,527	51079.602	297,809	436,115	***1,401,825***
2016	15,313	939.541	58,362	6937.857	No est.	62,215	39261.926	117,789	67,965	***368,784***

Due to these inconsistencies, the CSIR data for 2016 has been assumed to be valid in terms of proportions of the total employment by MesoZone and SIC category, but not in terms of absolute values. Therefore, the CSIR 2016 data have been factored to match the

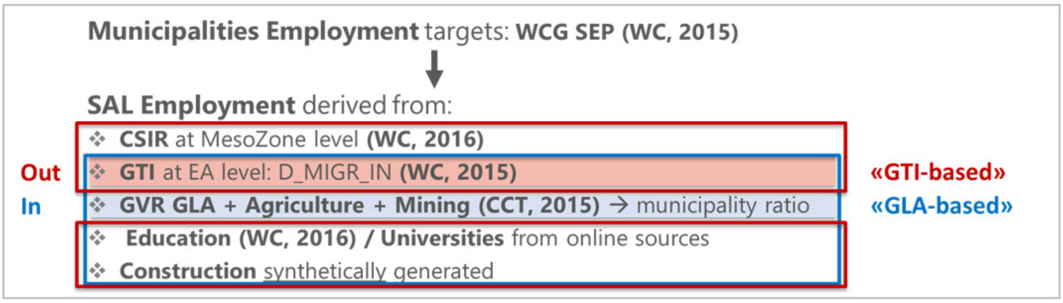
SEP totals by municipality and have then been used, together with the EA GTI data, to calculate SAL employment **outside CCT**.

3.1.4 GVR GLA 2015 (SAL – CCT)

General Valuation Rolls (GVR) data provides information on each “portion of building” **within CCT** at Land Parcel (LP) level. GVR data have been processed to obtain Gross Leasable Area (GLA) in square meters, for the non-residential establishments, Building/Habitable Floor Area (BFA/HFA) in square meters, for the residential establishments, Parking Areas (places where vehicles are parked) in square meters, aggregated at SAL level.

3.2 Employment Data Estimation Method

Figure 3 summarises the data sources used to derive the employment data by SAL for the Western Cape province, differently outside and inside the City of Cape Town.



Municipalities Employment targets: **WCG SEP (WC, 2015)**

SAL Employment derived from:

- **CSIR** at MesoZone level **(WC, 2016)**
- **GTI** at EA level: D_MIGR_IN **(WC, 2015)**
- **GVR GLA + Agriculture + Mining (CCT, 2015)** → municipality ratio
- **Education (WC, 2016) / Universities** from online sources
- **Construction** synthetically generated

Source: Bentley Systems

Figure 3: Employment data: schematic representation of used data-sources

3.2.1 “GTI-based”: Employment Data Estimation Outside the City of Cape Town

The GTI data is used to estimate employment for the entire WC province. This dataset is used directly only outside CCT (“GTI-based”), because inside CCT it is only used to undertake some local corrections to the “GLA-based” estimates. The “GTI-based” estimation is undertaken at EA level and aggregated at SAL level only as final step.

Education and Universities employment estimations have been treated separately and integrated with the above source, whilst construction employment has been synthetically generated. This was done due to the difficulty of identifying exactly where construction workers work on any given day. A process was created to impute construction employment by zone, as described below.

The **D_MIGR_IN** attribute provided by GTI has been used as the reference number of jobs at EA level. The first step in the process was to factor D_MIGR_IN to match the 2016 targets municipalities totals (SEP). The factored D_MIGR_IN attribute at EA level represents the total employment within the EA area, but it does not provide information on the employment sector. The steps below have been undertaken to disaggregate the employment values into the economic categories at EA level consistently with the other data sources.

To be able to use the MesoZones proportions for the SIC employment categories for estimating the segmentation of the GTI data, it is first necessary to establish the correspondence between MesoZone and EA. Because the boundaries of these layers do not align, a geoprocessing operation is necessary to calculate the “overlapping area” percentages.

The GTI EA land use data (agricultural, mining, industrial, other) are used, calculating the proportion of the different land uses for every “overlapping area”.

The combination of these proportions is used to obtain the estimated employment at EA level, consistent with the GTI data. These values are then factored by the municipality totals ratios and bucket rounding is applied. Note that these values do not include the construction sector.

While undertaking the process described above, several inconsistencies were detected in the data, in terms of land use segmentation for GTI EAs and employment proportions for CSIR MesoZones. Nevertheless, no other data were available to obtain the required segmentation, therefore the level of approximation has been considered acceptable and manual corrections have been entered for those zones where problems could be detected.

Construction jobs are very dynamic and very difficult to be spatialised. Nevertheless, they represent a relatively large proportion of jobs across the province (around 8%). For this reason, it is necessary to estimate their distribution at the same level of geography used for the other sectors (EA). To do this, a synthetic distribution has been implemented as a function of available land, value of real estate, and existing residential and employment density.

SEP provides the total number of workers in the construction sector to be around 197 481 employees in WC. This was our starting point, and construction workers by TAZ are estimated as a function of available land, value of real estate, and existing residential and employment density. Bucket rounding is then applied to make sure that the total employment matches the target totals from SEP at the municipality level, obtaining: SIC_5_GTI (EA)

Because mining employment represents a small proportion of employment in many municipalities with a provincial total of 1,577 employees (SEP), the estimation and rounding mechanisms created an overall mismatch between the estimated and target (SEP) municipality totals for SIC_2. Therefore, the municipality totals have been matched by correcting those EAs with inconsistencies: SIC_2_GTI (EA)

The WCG Department of Education provided disaggregated data for the education sector (schools, not including universities), where one record corresponds to one educational facility, including number of learners and number of staff working in the facility. These data have been aggregated at the EA and SAL levels, for integration into the demographics and employment database. The process also checks that the *Education Employment* must be higher than zero where the number of learners have a positive value in the EA. Then, the education employment from this data source has been used to replace the value of employment in the SIC_9 sector: SIC_9_GTI (EA).

University employment has been treated separately, with a specific estimation at the local level. In particular, assumptions were adopted for the estimation or adjustment of the employment values for the corresponding EAs for SIC_8 (Finance, Insurance, Real Estate and Business Services) and SIC_9 (Community, Social and Personal Services, Public Administration including General Government).

The last step of the process, before the final adjustments to municipality totals (SEP targets), consists of “manually” checking single zones with anomalies (e.g. zero or very large employment numbers, etc.) and imposing corrections through a table where values were manually adjusted (mostly based on observation of satellite imagery and street-view photos or estimations based on similar establishments). Then, the final step was to adjust and round the EAs employment values to match the municipality totals (SEP targets) with factors by municipality and SIC category.

3.2.2 "GLA-Based": Employment Data Estimation Inside the City of Cape Town

Inside CCT, the main source of data used for the estimation of the employment is given by the GVR GLA data ("GLA-based"), combined with the Agriculture and Mining areas, at SAL level.

Education and Universities employment estimations have been treated separately and integrated with the above source, whilst construction employment has been synthetically generated, as noted above.

The processing of the GVR data on each "portion of building" within CCT followed the steps below:

1. Analysis of DOR (overall Land Parcel use code and description for residential and non-residential buildings) and SAR (building Section Use code and description).
2. Definition of "intermediate categories" based on SAR and DOR cross-classification.
3. Aggregation of "portion of building" data into "building" data with "intermediate categories" – calculating GLA, HFA, BFA and Parking Area, together with counts of single "portion of buildings" contributing to the calculation of areas.
4. Definition of Land Parcel – SAL correspondence table from multiple sources (filling holes for SAL and due to non-unique definition of Land Parcels).
5. Aggregation of the GLA and HFA, BFA and Parking Area from Land Parcel to SAL level.

Agriculture (2017) and mining (2013-2014) data have been geoprocessed and aggregated at the SAL level based on the spatial data obtained from the WCG departments and ESRI SA. This allowed us to obtain GLA for agriculture and mining in square meters, to be integrated with GVR data for the other land uses.

From GVR, Agriculture and Mining data, the GLAs for CCT are obtained, together with the conversion percentages for allocation from "intermediate categories" to SIC employment categories. From these values and conversion percentages, the GLA for the SIC categories have been calculated at the municipality level and at the SAL level. These categories correspond to the employment categories from SEP in Table 1. This allowed us to calculate the employment / GLA ratios to determine the employment from GLAs at SAL level within CCT. After applying the ratios to derive employment at SAL level, these numbers were rounded and factored to match the target municipality totals. No construction employment values can be derived with this method, due to unavailability of construction GLA.

The same approach used outside CCT has been employed inside CCT to estimate construction jobs.

A correction to the estimated employment inside of CCT was undertaken by using the estimates from the "GTI-based" approach. The employment total targets (SEP) implicitly include formal and informal sectors, nevertheless, the distribution at SAL level of the informal sector for manufacturing, trade and transport was adjusted, due to missing GLA values for the informal sector. This is done by using the EA estimated data for those sectors accordingly to the following criteria: if the total “GLA-based” employment is 0 but the total “GTI-based” is higher than zero, then, SIC_3, SIC_6 and SIC_7 are obtained from the “GTI-based” employment.

The same approach used outside CCT has been employed inside CCT to estimate correct education/university employment.

The last step of the process, before the final adjustments to municipality totals, consists of checking single zones with anomalies and imposing corrections in a similar way to what done outside CCT. Then, the final step was to adjust and round the SLAs employment values to match the municipality totals (SEP targets).

4. DEMOGRAPHIC DATA

4.1 Demographic Data Needs

Demographic data are generally updated, verified and accessible from official sources, such as the national statistical department, providing household and population data consistently in a centralised platform. This is not the case in Western Cape, and different data sources needed to be analysed with internal inconsistencies.

In particular, the following information is necessary for the development of the WCLUTI model:

- Household data for the reference base year 2016 at the SAL level, including:
  - Total HHs by SAL,
  - HHs stratification and “cross-tabulation” by SAL.
- Population data for the reference base year 2016 at the SAL level, including:
  - Total Population by SAL.
- Average HH income by SAL for the reference base year 2016.

Population includes all resident population, regardless of age. The usual definition of *household* is an occupied dwelling unit. If multiple families live in one structure, that is still considered one “household”. *Dwelling unit* includes all forms of habitable structure, including single-family housing, multi-family housing, “informal” housing, and the like. For each zone, the average HH size must be 1.0 or greater. Unoccupied dwelling units should not be included.

Another key population-related data item is household income. The average zonal income is a driver of location choices for other HHs. Moreover, income is a major determinant of the number of trips a HH makes, since a high proportion of trips are made either to make money or to spend it. In addition, income is strongly related to vehicle ownership and thus to the choice of travel mode. Although accurate income data is difficult to obtain, due to the reluctance of many people to disclose it in surveys, it is important enough that estimated data, or data without high confidence, is still a valuable indicator of travel habits.

These data are used both for the estimation/calibration of parts of the models, and as input for the model to run in the base year.

4.2 Demographic Data Sources

The main sources that have been used are summarised here below.

- Aggregate demographic data at some level of spatial aggregation:
  - **STATS SA Community Survey (CS) 2016**, providing HH and Population totals at ***municipality*** level (WC, 2016).
  - **STATS SA Census 2011**, providing HH and Population at ***SAL*** level (WC, 2011).
  - **GTI 2015**, providing HH and Population at ***EA*** level (WC, 2015).
  - **CCT 2015**, providing estimates for HH at CCT TAZ level (CCT, 2015) – not used.

- Auxiliary data on Informal Residential Locations provided by the City of Cape Town (CCT) and the WC Department of Human Settlements (DHS):
  - CCT backyard dwelling points for CCT (CCT no WC, 2013) – Note: no backyard dwellings data available outside CCT.
  - DHS and CCT Informal Settlements Pocket Areas for CCT (CCT, 2011 from DHS and 2013 from CCT) – Note: only the 2013 data used, as more recent data source.
  - DHS Informal Settlements Pocket Areas for WC outside CCT, containing an estimation of the population (WC no CCT, 2016)

4.2.1 STATS SA Community Survey 2016 (Municipality – WC)

"Provincial Profile Western Cape" – Report 03-01-07 and "Provinces at a Glance" – Report 03-01-03 provide the municipality totals for 2016 HH and Population data, considered the targets to be matched at the aggregated level (Table 3).

Table 3: HH and Population from STATS SA Community Survey 2016 – "targets" for 2016

Area	**2016 POPULATION**	**2016 HOUSEHOLDS**
City of Cape Town (municipality)	4,004,793	1,264,849
Western Cape Province	**6,279,730**	**1,933,876**

4.2.2 STATS SA Census 2011 (SAL – WC)

Census 2011 is the main source of information for every attribute related to Households and for their "cross-classification" (number of HHs belonging to categories at multiple level, e.g., HHs by population group and by income) at the SAL level.

The residential location and supply models of the WCLM have been estimated with Census 2011 data at the Small Area Layer (SAL) level of aggregation, thanks to the possibility of cross-tabulating attributes using "superCROSS" software (even if this is possible only for a sub-selection of the entire dataset). No other data was available in WC, providing this type of cross-tabulated information, allowing the analysis of both HH characteristics and location patterns (i.e., dwelling type and SAL zone). Whilst the HHs stratification and "cross-tabulation" by SAL is fundamental for the formulation and estimation of the model, it will not be the focus of this section, that will mainly cover the issues and methodology adopted to obtain total values of HHs and Population at SAL level.

Several issues have been encountered while processing Census 2011 data, the most relevant are discussed below.

The two most important issues related to HH and Population totals:

- Census 2011 data have an internal inconsistency issue, which seems possibly related to the rounding mechanism used for small geographic areas. Extracting data for different HH attributes results in different HH totals by SAL and different overall totals for the province, depending on the type of data extracted. This does not provide consistent numbers between data extractions.

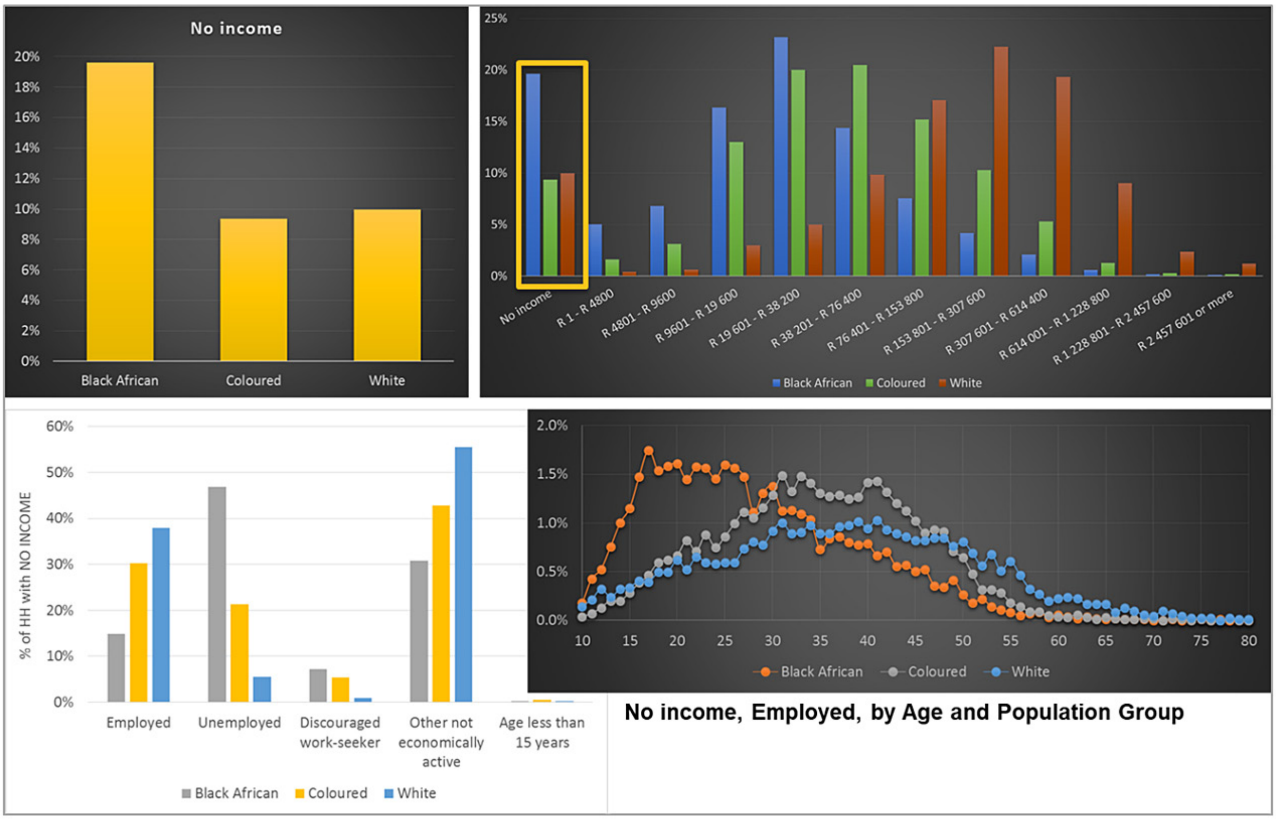
- The Population data from Census 2011 (POP_CENS2011) may have anomalies for the following reasons:
  - There are SALs with 0 households, but with POP_CENS11 > 0.
  - There are multiple SALs where POP_CENS11 is just repeating across the records (multiple SALs with same value).
  - Numbers are not consistent with the estimation of population from the HH size categories or from any other data source.

To overcome the first limitation, the HH totals from Census 2011 have been aggregated by SAL from the number of HHs extracted using “superCROSS” software, with a cross-tabulation of the following attributes: HH size, Income, Population Group, Dwelling Type and Number of Rooms. This ensures consistency with the location patterns that are analysed for the estimation of the Western Cape Land Model (WCLM).

To overcome the second limitation, the population by SAL from Census 2011 data has been estimated from the extraction of number of HHs stratified by HH size, multiplying the number of HHs in each category times the average size of the HH for the category.

Census 2011 data provides the number of HHs disaggregated by dwelling type, with twelve categories. The informal components can be isolated from the dataset in terms of Backyard Dwellings (Category 8) and Informal Settlements (Category 9 + Category 11). Comparing the number of HHs in these categories with the available auxiliary data, large discrepancies can be identified with an underestimation of informal HHs from Census. This could be due to different time scales of data collection and the dynamic nature of these dwellings, but also, for backyard dwellings, the Census data was apparently including units as addition to the formal housing counts.

Referring to Income data, the large “No-income” category for HH income data is clearly due to a missing response option in the Census survey for “not-declared/no-specified” income. Figure 4 shows that this category is not actually representing “zero-income” indeed. This “No-income” HHs represent nearly the 14% of income distribution and has been incorporated to the dataset by imputing their incomes. The imputation process consists of grouping HHs, based on available attributes that are showing to be highly correlated with income and defining the income level ensuring that the distribution is the same as the one of the HHs in the same group.



Source: Stats SA

Figure 4: No-Income category from Census 2011 data

4.2.3 GTI 2015 (EA – WC)

The GTI "*demographics*" dataset is based on 2011 Census information and updated with the 2014 mid-year estimates released by STATS SA. The data is statistically modelled using the 2012 building based land use points, and 2016 land use polygons where no points are available at the EA level, providing:

- The night-time Population figures given in the **POPULATION** field in the Population dataset.
- The total HHs figures given in the **TOTAL_HH** field in the Demographics dataset.

In line with what was observed and reported for the Employment data, inconsistencies were found within this data, particularly within the City of Cape Town (e.g., one data source identifying zero households in a zone, whilst the other source identifying several households in the same zone, etc.).

The TOTAL_HH and POPULATION numbers have been factored by municipality to match the 2016 targets. These factored values have been assumed to provide HHs and Population distributions for the WC province.

The "*residential activity*" attributes in the dataset provide the area in ha within the EA covered by formal residential vs informal residential (**FORMAL_RES**, **INFORM_RES**). By analysing this data provided by GTI and comparing them with the auxiliary data, it was not possible to detect valid and consistent patterns, therefore these attributes have been discarded and not used.

The "Lifestyle" attributes in the database provide the median and average income by EA with the attributes **MED_IN** and **A_IND_IN** respectively. According to available documentation, these values should have been derived from the Census 2011 income categories weighted average.

4.3 Demographic Data Estimation Method

As reported above, an official, cleaned and verified single source of demographic data at SAL level was not available, therefore this project's dataset has been estimated from the above sources with the following methodology:

- Analysis of existing data sources and selection of more reliable / updated source as the basis for the estimation of HHs and Population:
  - Outside CCT: GTI 2015 ("GTI-based").
  - Inside CCT: Census 2011 ("Census-based") – population estimated from HH size.
- Integration of informal settlements and backyard dwellings data from auxiliary sources: few adjustments to Pocket Areas data outside CCT from DHS were necessary, and all these data have been summarised at SAL level (geoprocessing).
- Manual estimation of HHs and Population for the refinement of specific zones with incorrect values.
- Factoring of the estimated HHs and Population, to match the targets HHs and Population totals by municipality (Community Survey 2016).
- Integration of University Housing students (assumption 1 HH = 1 student) – students residing in University Housing in the main University Campuses have been considered on top of the municipality targets.
- Estimation of the average income per HH by SAL:
  - Income imputation to correct the 0-income category issue.
  - Average income estimation from Census 2011.
  - Correction of missing income information.

5. CONCLUSIONS

The lack of socio-economic data consistency in the WC province at a sufficient level of disaggregation was driving the task of analysing and processing multiple data sources to create a reference dataset for the 2016 Base Year at the Small Area Layer (SAL) level of aggregation.

The adopted methodology was focussing on assembling existing original employment, demographics and land use data and applying manual corrections through local observations. This allowed the creation of a dataset that could be considered acceptable for the development and calibration of the 2016 Base Year scenario of the WCLUTI Model.

As new information becomes available, earlier data is made more detailed and accurate. Therefore, this is a continuous activity that will be built into the model development process.

6. REFERENCES

WCLUTI: Western Cape Land Use & Transport Interaction Model, Documentation & Reference Guide, Version 1.1, prepared for Western Cape Government by Bentley Systems and Esri South Africa, December 2019 (unpublished).

2013-2014 South African National Land-Cover Dataset, Data User Report and MetaData, GeoTerra Image SA, February 2015.