Guide to version 1 of the Police Precinct Crime & Demographic Data 2011

Amy Thornton

May 2026

The Police Precinct Crime and Demographic (PPCD) Data, 2011, is a police-precinct-level dataset that comprises just over 1000 police precincts with almost exhaustive geographic coverage of South Africa with both crime records and socio-economic and demographic information for each precinct. We combine the South African Police Service's (SAPS) official 2011 Crime Statistics (SAPS, 2015) and the 2011 South African Census Community Profiles (Statistics South Africa, 2011) which is the Small Area Layer data from the 2011 South African Census collected by Statistics South Africa (StatsSA). The SAPS and the Community Profiles datasets are publicly available from DataFirst. The SAPS Crime Statistics provide the crime reported in each police precinct, grouped into 27 crime categories. The Census Community Profiles comprise demographic information from the census data aggregated to the small area layer (SAL) level. SALs are census enumeration areas including at least 500 people.

To combine the datasets, each of the 85 000 SALs from the Census Community Profiles Data are allocated to one of the 1 124 police precincts according to their geographic boundaries, through geospatial mapping done by Tim Brophy, now at the SALDRU Survey & Data Hub. He utilised the SAL boundary geographic information system (GIS) data and generated a random point in the polygon algorithm that fell within each SAL's boundary. These were then mapped to the 2015 police station boundary data. Using a random point does not guarantee that the majority of the SAL would be in the same precinct as the random point itself. However, given that the irregular shape of the SALs can often lead to a central point or centroid falling completely outside of the SALs boundary, the random point minimises the potential error.

Thus, every South African in the 2011 census was allocated to a police precinct based on the SAL they were surveyed in, and the SAL-level demographic and socio-economic data was then aggregated up to the police precinct level and merged with the SAPS station crime statistics for 2011. The result is a data set of 1 124 police precincts covering the 52 million South Africans surveyed in the 2011 census.[1] There are 39 different types of crime records from 2011 and from the previous five years and 67 different demographic and socio-economic variables.

The dataset was created as part of a project at the Development Policy Research Unit (DPRU) at the University of Cape Town funded by the World Bank that started in 2017. The project eventually produced a journal article by Thornton et al., (2022) about how burglary rates vary with precinct-level income and inequality. More information about how the dataset was created and specific variables is in this journal article although we also provide a detailed variable list below with relevant data quality notes and release the syntax to create the dataset.

The data are released in two datasets, one at the precinct level (n=1152) and one in which the precincts in the eight main metropolitan areas are merged together since precincts get very small in terms of square kilometre area within metros (n=885). This structure somewhat reflects the trajectory of the research process behind the journal article. The data can be cited as follows: **Thornton, A., Bhorat, H., Lilenstein, A., Monnakgotla, M. and van der Zee, K. Police Precinct Crime & Demographics 2011 [dataset].**

[1] This only excludes 14 police precincts that existed in the 2015 SAPS geographic boundaries, whose police stations were only built after 2011, so they did not exist in the 2011 crime statistics. We also exclude the precinct of O.R. Tambo International Airport from the dataset because it stands out as an outlier in terms of crime rates and numerous descriptive statistics.

Version 1. Cape Town: DataFirst, DPRU, ARC and REEP [producers], 2025. Cape Town: DataFirst [distributor], 2026. DOI: https://doi.org/10.25828/8bgt-1578

The datasets are not released with any form of post-stratified weights. Although the Community Profiles Data captures everyone who was captured by census enumeration, there is no adjustment for those who were missed by census enumeration. Normally, official census numbers are some combination of the enumerated census population and adjustment factors for non-response and coverage error based on the Post-Enumeration Survey. No such adjustment factors are in this dataset.

1 Variable list and description

A list and description of all variables in the dataset, grouped by topic.

1.1 Identifiers and geography

police_district precinct name; the primary unit of observation in ppcd-2011-v1.dta. Unique identifier within the dataset. There are 1 152 unique values.

pr_code province code of the precinct. Derived as a collapsed median of the SAL-to-province mapping. Range = 1–9, corresponding to South Africa's nine provinces.

police_district_mm precinct or metro identifier in ppcd-metros-2011-v1.dta. Copy of police_district but with the high number of stations in South Africa's eight metropolitan areas aggregated to a single metro label per metro. There are 885 unique values.

1.2 Demographics and population base

pdcpop total population of the precinct derived from Census SAL records. Core denominator for per-capita crime rates and demographic share variables.

propmale proportion of the precinct population that is male. Constructed as male count divided by total population.

totpop sum of pdcpop across all precincts in the dataset which is the total population count for South Africa in the SAL data coming to 51.7 million.

1.3 Geo-type proportions

urbanprop share of the precinct population living in urban SALs.

tribalprop share of the precinct population living in tribal-area SALs.

farmprop share of the precinct population living in farm SALs.

1.4 Citizenship and household-head gender

propnoncitz share of the precinct population that is not a South African citizen. Not available in ppcd-metros-2011-v1.dta.

prophhhfemale share of households with a female household head. Not available in ppcd-metros-2011-v1.dta.

heads proportion of people in the precinct coded as household heads; rel1_pdc / totpdc. Not available in ppcd-metros-2011-v1.dta.

1.5 Age profile

avgage average age of the precinct population; population-weighted mean over single-year age bins.

youthlf population aged 15–30 as a share of precinct population aged 15–64 (youth share of labour-force-age population). Available in ppcd-2011-v1.dta only.

youth population aged 15–30 as a share of total precinct population. Available in ppcd-metros-2011-v1.dta only.

allyoung population aged ≤30 as a share of total precinct population. Available in ppcd-2011-v1.dta only.

allold population aged >64 as a share of total precinct population. Available in ppcd-2011-v1.dta only.

depend dependency ratio: population aged <15 or >64 as a share of total precinct population. Available in ppcd-2011-v1.dta only.

workingage population aged 15–64 as a share of total precinct population. Available in ppcd-2011-v1.dta only.

1.6 Income proxies and transforms

Note: the household income bracket variable used in many of the variables in this section suffers from a data quality issue. In the census 10% sample microdata, a very high share of 15% of households have zero household income (bracket 1). The household income variable in the census was constructed by summing the one-shot personal income question and when personal income was missing, this was treated as zero. About 12% of households have at least one person with unspecified household income in the census 10% sample microdata. Kerr et al. (2026) tried to address by imputing for those with missing personal income but only reduced the share of households with zero-income from 15% to 12.5%, suggesting there are other reasons why there is such a high share of zero-income households in census 2011 that is not fully understood. More about this data quality issue is in Kerr et al. (2026) and Yu (2016).

ainc1prop share of the precinct population in the lower income bracket set (brackets 1–4 of the Census annual household income question). Available in ppcd-2011-v1.dta only.

ainc2prop share of the precinct population in the middle-income bracket set (brackets 5–8 of the Census annual household income question). Available in ppcd-2011-v1.dta only.

ainc3prop share of the precinct population in the upper income bracket set (brackets 9–12 of the Census annual household income question). Available in ppcd-2011-v1.dta only.

avgincrank weighted average bracket rank index (bracket number used as income rank proxy). Available in ppcd-2011-v1.dta only.

avgincrank2 square of avgincrank.

avgpcinc annual per-capita income for the precinct. This was calculated in by taking counts of households per annual income bracket multiplied by bracket midpoints, summing to a total precinct household income, and then dividing by pdcpop.

avgpcinc2 square of avgpcinc.

lnavgpcinc natural log of avgpcinc.

lnavgpcinc2 square of lnavgpcinc.

incterc income tercile assignment from xtile avgpcinc, nq(3).

1.7 Household inequality measures

Note: all of these are measures of household income inequality and *not* the more commonly found per capita household income inequality. This is due to limitations regarding the aggregated SAL data. Whilst we know how many households fall into which income brackets and how many households are of what household size,

we cannot link household size to income bracket. A note also to check the data quality note in the section above on household income since this will have implications for the measures of household inequality in this section.

ge0 mean log deviation (GE(0) index); constructed using epsilon for the zero-income bracket.

ge1 Theil index (GE(1)); constructed using epsilon for the zero-income bracket.

ge2 GE(2) index; constructed using zero for the zero-income bracket in ppcd-2011-v1.dta and epsilon in ppcd-metros-2011-v1.dta.

gini Gini coefficient; constructed using zero for the zero-income bracket in ppcd-2011-v1.dta and epsilon in ppcd-metros-2011-v1.dta.

gini2 square of gini.

gini3 cube of gini.

atkhalf Atkinson inequality index at aversion parameter 0.5.

atk1 Atkinson inequality index at aversion parameter 1.

atk2 Atkinson inequality index at aversion parameter 2. Available in ppcd-2011-v1.dta only.

cvar coefficient of variation of the income distribution. Available in ppcd-2011-v1.dta only.

p9010 90/10 income ratio. Available in ppcd-2011-v1.dta only.

1.8 Language diversity

langind1–langind8, langind10–langind13 indicator variables equal to 1 if the corresponding first-language group is represented in the precinct. Available in ppcd-2011-v1.dta only.

numlang total number of first-language groups represented in the precinct; rowtotal(langind*). Available in ppcd-2011-v1.dta only.

1.9 Labour market status

empoftot share of total precinct population that is employed (official definition, empsta_1–empsta_6 denominator).

unempoftot strictly unemployed share of total precinct population.

discoftot discouraged work-seeker share of total precinct population.

neaoftot not economically active share of total precinct population.

bunempoftot broad unemployment share of total precinct population (strictly unemployed plus discouraged).

empoftot2 square of empoftot.

empoflf precinct employment rate using labour force (employed + strictly unemployed) as denominator.

unempoflf strict precinct unemployment rate using labour force as denominator.

unempoflf2 square of unempoflf.

discoflf discouraged work-seeker share of the precinct labour force.

bunempoflf precinct broad unemployment rate using labour force as denominator.

1.10 Household structure

hhpdc number of households in the precinct. Not available in ppcd-metros-2011-v1.dta.

avghhsize average household size; pdcpop / hhpdc. Not available in ppcd-metros-2011-v1.dta.

tothh total household count for the country across all precincts. This was 14.4 million. Not available in ppcd-metros-2011-v1.dta.

avgrooms average number of rooms per household. Not available in ppcd-metros-2011-v1.dta.

1.11 Population group shares

africanprop share of the precinct population identifying as African/Black.

colouredprop share of the precinct population identifying as Coloured.

asianprop share of the precinct population identifying as Indian/Asian.

whiteprop share of the precinct population identifying as White.

otherraceprop share of the precinct population in other population groups.

1.12 Education profile

educ1prop share of population with no schooling.

educ2prop share of population with some primary education.

educ3prop share of population with completed primary education.

educ4prop share of population with some secondary education.

educ5prop share of population who completed Grade 12 (matriculated).

educ6prop share of population with higher education.

matricplus share of population with Grade 12 or higher; educ5prop + educ6prop. Available in ppcd-metros-2011-v1.dta only.

1.13 Location and land-use type

location dominant area type for the precinct; coded as the enumeration area type with >50% share, or 11 (Mixed) if no single type dominates. Not available in ppcd-metros-2011-v1.dta.

ecoact combined commercial and industrial enumeration area share; proxy for economic activity (type10 + type7). Not available in ppcd-metros-2011-v1.dta.

farmy combined farm and small-holding enumeration area share (type4 + type8). Not available in ppcd-metros-2011-v1.dta.

1.14 Crime base and categories

crime1–crime38, crime47 base crime counts from cleaned SAPS records. One variable per SAPS crime category retained after cleaning. SAPS aggregate categories crime39–crime46 are dropped. Crime categories are labelled in the dataset according to the original SAPS crime statistics excel spreadsheet.

totcr total crime count across all retained base categories; rowtotal(crime*). Stations with totcr == 0 are dropped.

detpol count of crime detected as a result of police action: crime19 + crime18 + crime20 + crime34. Defined as per SAPS Crime Report 2010/11 although with the addition of crime34 (Sexual offences detected as result of police action).

robbery common robbery and aggravated robbery (excl. crime32): crime14 + crime30 + crime29 + crime11 + crime47 + crime31 + crime8. Defined similarly to combined categories of robbery in SAPS Crime Report 2010/11.

property property crime count: crime9 + crime10 + crime37 + crime38 + crime36. Defined as per SAPS Crime Report 2010/11 property-related crime category.

murder murder count: crime23.

sexualcr sexual offences count: crime7 + crime15 + crime26 + crime28 + crime33. Own definition of combining all sexual-related crimes.

assault assault count: crime5 + crime13 + crime25. Own definition combining all assault-related crimes.

{crimecat}pcrate precinct per-person rate for category {crimecat}; {crimecat} / pdcpop.

{crimecat}krate precinct per-100,000 rate for category {crimecat}; ({crimecat} / pdcpop) × 100,000.

l{crimecat}pcrate natural log of {crimecat}pcrate.

l{crimecat}krate natural log of {crimecat}krate.

burglary burglary count (crime10); in ppcd-metros-2011-v1.dta.

1.15 Five-year prior crime panel (2006–2010)

The following variables are five-year averages over 2006–2010 from previous_five_years_crime.dta. Rate and log forms follow the same naming convention as the 2011/12 variables with the addition of the _0610 suffix.

detpol_0610 average annual detection-oriented policing count, 2006–2010.

detpolkrate_0610 average per-100,000 detection-oriented policing rate, 2006–2010.

ldetpolkrate_0610 log of detpolkrate_0610.

robbery_0610 average annual robbery count, 2006–2010.

robberykrate_0610 average per-100,000 robbery rate, 2006–2010.

lrobberykrate_0610 log of robberykrate_0610.

property_0610 average annual property crime count, 2006–2010.

propertykrate_0610 average per-100,000 property crime rate, 2006–2010.

lpropertykrate_0610 log of propertykrate_0610.

murder_0610 average annual murder count, 2006–2010.

murderkrate_0610 average per-100,000 murder rate, 2006–2010.

lmurderkrate_0610 log of murderkrate_0610.

sexualcr_0610 average annual sexual offences count, 2006–2010.

sexualcrkrate_0610 average per-100,000 sexual offences rate, 2006–2010.

lsexualcrkrate_0610 log of sexualcrkrate_0610.

assault_0610 average annual assault count, 2006–2010.

assaultkrate_0610 average per-100,000 assault rate, 2006–2010.

lassaultkrate_0610 log of assaultkrate_0610.

1.16 Province SAPS resource variables

All province SAPS variables take a single value for all precincts within a province. They were taken from the SAPS Annual Report and are not station-specific measurements. Not available in ppcd-metros-2011-v1.dta.

provstaff total SAPS personnel in the province.

provvehicle total SAPS vehicles in the province.

provbulletproofvest bulletproof vests available in the province.

provfirearms firearms available in the province.

provvicsupport victim support centres in the province.

provaccrinstitute accreditation institutes in the province.

provescape escapes from SAPS custody recorded in the province.

provsapsmurder SAPS murder rate for the province.

provreacttimealpha average reaction time for alpha-priority calls in the province.

provreacttimebravo average reaction time for bravo-priority calls in the province.

provreacttimecharlie average reaction time for charlie-priority calls in the province.

provgunrecovery firearms recovered in the province.

provgunsurrender firearms surrendered in the province.

provammusurrender ammunition surrendered in the province.

provvehrecoverprop proportion of stolen vehicles recovered in the province.

provpop sum of pdcpop across precincts in the same province; used to compute province-normalised ratios.

motor_ratio vehicles per capita in the province: provvehicle / provpop.

staff_ratio SAPS personnel per capita in the province: provstaff / provpop.

vicrm_ratio victim support centres per capita in the province: provvicsupport / provpop.

1.17 Spatial variables

area precinct area in square kilometres, from station boundary files. Available in ppcd-2011-v1.dta.

areamm precinct area in square kilometres in the metro dataset. Available in ppcd-metros-2011-v1.dta only.

areamm_1000 areamm expressed in thousands of square kilometres. Available in ppcd-metros-2011-v1.dta only.

LOCATION_X longitude coordinate of a station for that precinct from the SAPS station boundaries data. For metros that aggregate many precincts, the longitude reflects the most central station. This was as follows for each metro: City of Cape Town (Cape Town Central station), City of Johannesburg (Johannesburg Central station), City of Tshwane (Pretoria Central station), Ekurhuleni (Boksburg station), eThekwini (Durban Central station), Buffalo City (East London station), Nelson Mandela Bay (Mount Road station), and Mangaung (Mangaung station).

LOCATION_Y latitude coordinate of a station for that precinct from the SAPS station boundaries data. For metros that aggregate many precincts, the longitude reflects the most central station. This was as follows for each metro: City of Cape Town (Cape Town Central station), City of Johannesburg (Johannesburg Central station), City of Tshwane (Pretoria Central station), Ekurhuleni (Boksburg station), eThekwini (Durban Central station), Buffalo City (East London station), Nelson Mandela Bay (Mount Road station), and Mangaung (Mangaung station).

2 Appendix C: Comparability checklist

Use this checklist before pooling or comparing models using ppcd-2011-v1.dta and ppcd-metros-2011-v1.dta.

Comparable with caution (same broad concept, possible formula differences):

- propmale, pdcpop, totpop, pr_code
- urbanprop, tribalprop, farmprop
- avgage
- avgpcinc, lnavgpcinc, lnavgpcinc2
- gini, ge0, ge1, ge2
- africanprop, colouredprop, asianprop, whiteprop, otherraceprop
- Labour status shares: empoftot, unempoftot, discoftot, neaoftot, bunempoftot, empoflf, unempoflf, discoflf, bunempoflf

Not directly comparable or absent in metro file (precinct-only):

- propnoncitz, prophhhfemale
- youthlf, allyoung, allold, depend, workingage
- ainc1prop, ainc2prop, ainc3prop
- numlang, langind*
- hhpdc, avghhsize, tothh
- avgrooms, heads
- atkhalf, atk1, atk2, cvar, p9010
- Most citizen-crime family variables beyond propertycr and burglary
- _0610 prior-period panel variables
- prov* variables and province ratio fields

Metro-only additions:

- youth, matricplus, areamm, areamm_1000, LOCATION_X, LOCATION_Y

References

Kerr, A., Thornton, A. and Barnard, C. (2026) Counting poor households fairly: the influence of data quality on the South African Local Government Equitable Share Formula. Policy Brief prepared for the Equality Collective. Available: https://www.equalitycollective.org.za/publications

SAPS (2015) Crime Statistics 2014-2015. South African Police Service (SAPS), Government of South Africa. Spreadsheet downloaded from https://www.saps.gov.za/

Statistics South Africa. South African Census Community Profiles 2011 [dataset]. Version 1. Pretoria: Statistics SA [producer], 2014. Cape Town: DataFirst [distributor], 2015. DOI: https://doi.org/10.25828/6n0m-7m52

Thornton, A., Bhorat, B., Lilenstein, A., Monnakgotla, J. and van der Zee, K. (2022) Crime, income and inequality: non-linearities under extreme inequality in South Africa. Economic Development and Cultural Change. Vol. 72(1). DOI: 10.1086/719646.

Yu, Derek, "Factors influencing the comparability of poverty estimates across household surveys," Development Southern Africa, 2016, 33 (2), 145–165.