Identifying Slum Areas in Johannesburg, South Africa

MICHAEL ONO, NEIL HAMLETT, DYLAN BLOUGH

GA DSI DC-10

Overview of Johannesburg

- Population of 4,434,827 people.
- Most populous city in South Africa
- 40 different sub-localities within the metro area.

://www.statssa.gov.za/?page_id=993&id=city-of-johannesburg-municipality

What is a slum?

formal settlements (*aka* "slums") can be defined as " highly opulated urban residential area consisting mostly of closely acked, decrepit housing units in a situation of deteriorated or complete infrastructure, inhabited primarily by impoverished ersons"

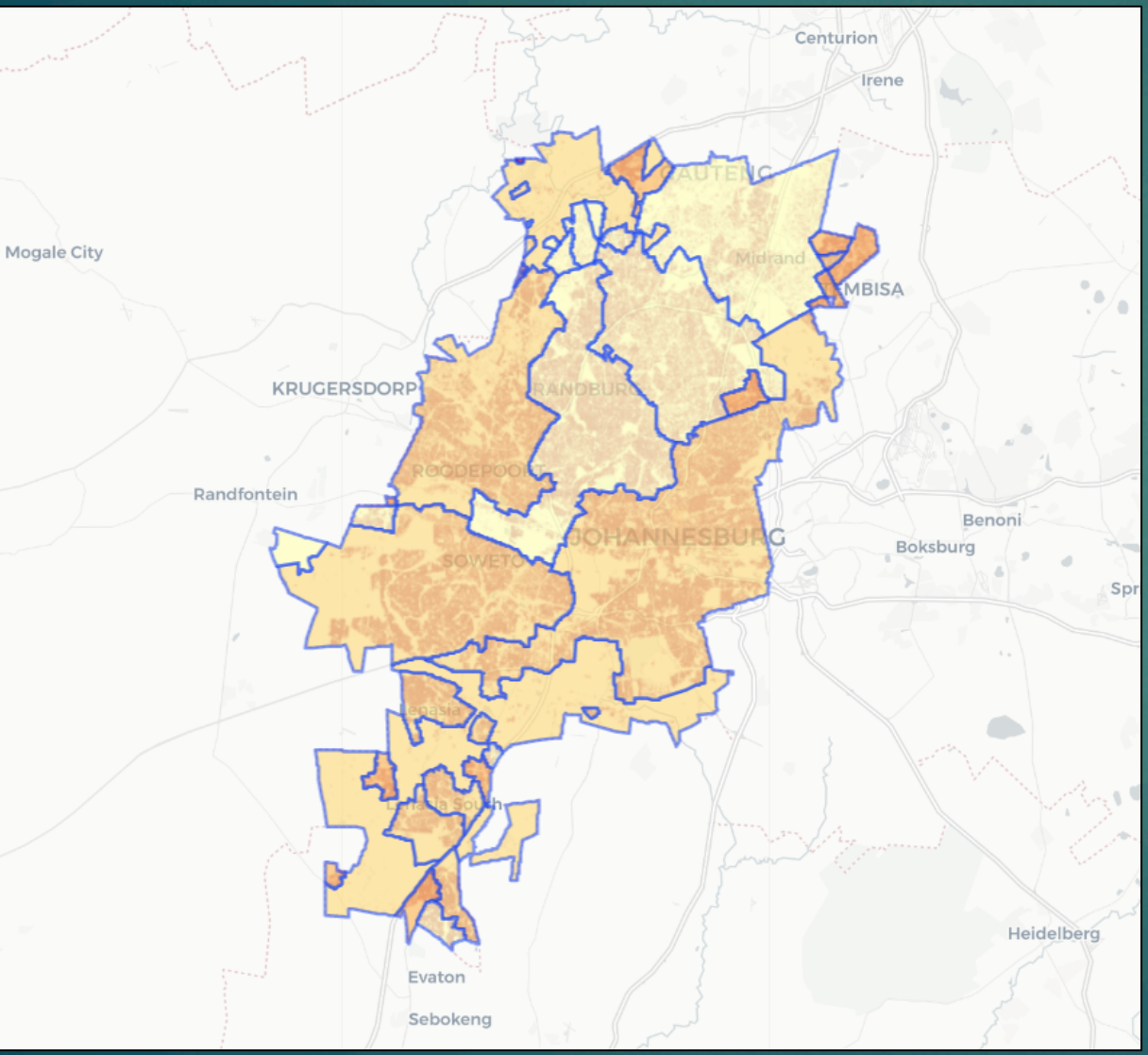
or the purposes of our project, *informal settlements* are urban sidential concentrations that are not supported by public rvices. Beyond infrastructure services (plumbing, electricity), ese also include legal status, such as property deeds.

roblem Statement & Data Sources

- We are seeking to identify local pop. concentrations that are not reflected in official (census) or commercial (real estate) records. Applying a *data-science* approach, we can attempt to compare unofficial measurements of local population densities with official or commercial sources. Our "slums" will be localities in which high concentrations of residents are indicated that do not coincide with "official" sources.
- **Data Sources**
  - **South African Census:** conducted in 2011, provides an overview of *legally sanctioned* residences
  - Facebook population density data

ttps://dataforgood.fb.com/

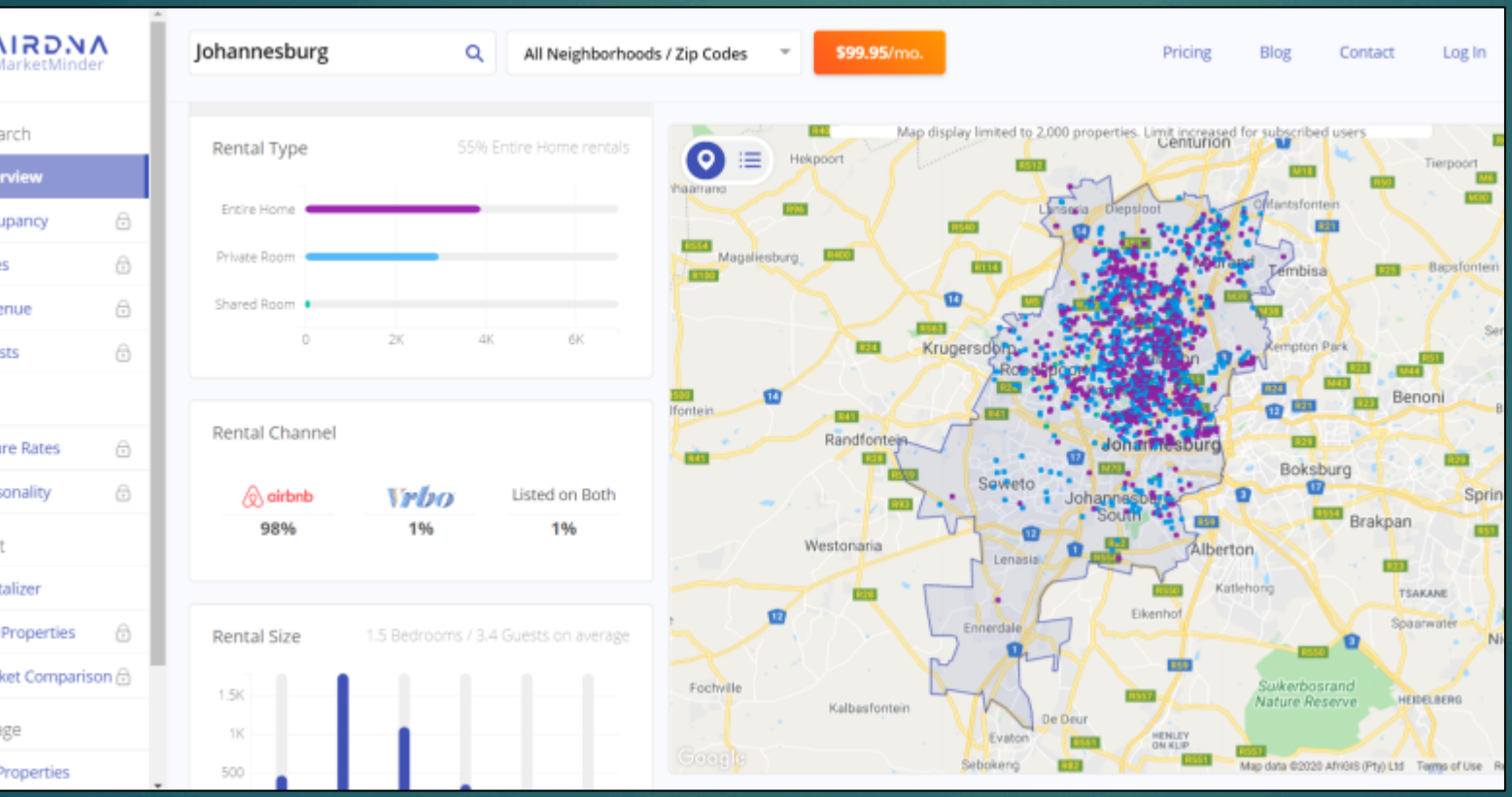
011 Census Acquired Data



Data Dictionary for Leaflet Interactive Map

Columns	Data Type
City	obj
Total Population	int
Young (0-14)	float
Working Age (15-64)	float
Elderly (65+)	float
Population density	obj
No schooling aged 20+	float
Higher education aged 20+	float
Matric aged 20+	float
Number of households	float
Average household size	float
Female headed households	float
Housing owned/paying off	float
Flush toilet connected to sewerage	float
Weekly refuse removal	float
Electricity for lighting	float
Percent No Income	float

ttempts at Collecting Real Estate ata For Johannesburg



Why use open-source data?

- Censuses occur infrequently and often using inconsistent methodology
- Records data at a low resolution with discrete distributions so we assume constant distribution throughout the census-designated administrative zones.
- By definition slums are underserved by government resources and making it harder to accurately gauge the population
  - Also important for slum identification: censuses not account for rapidly shifting populations (economic migrants, refugees etc.)

acebook's DataForGood in South frica

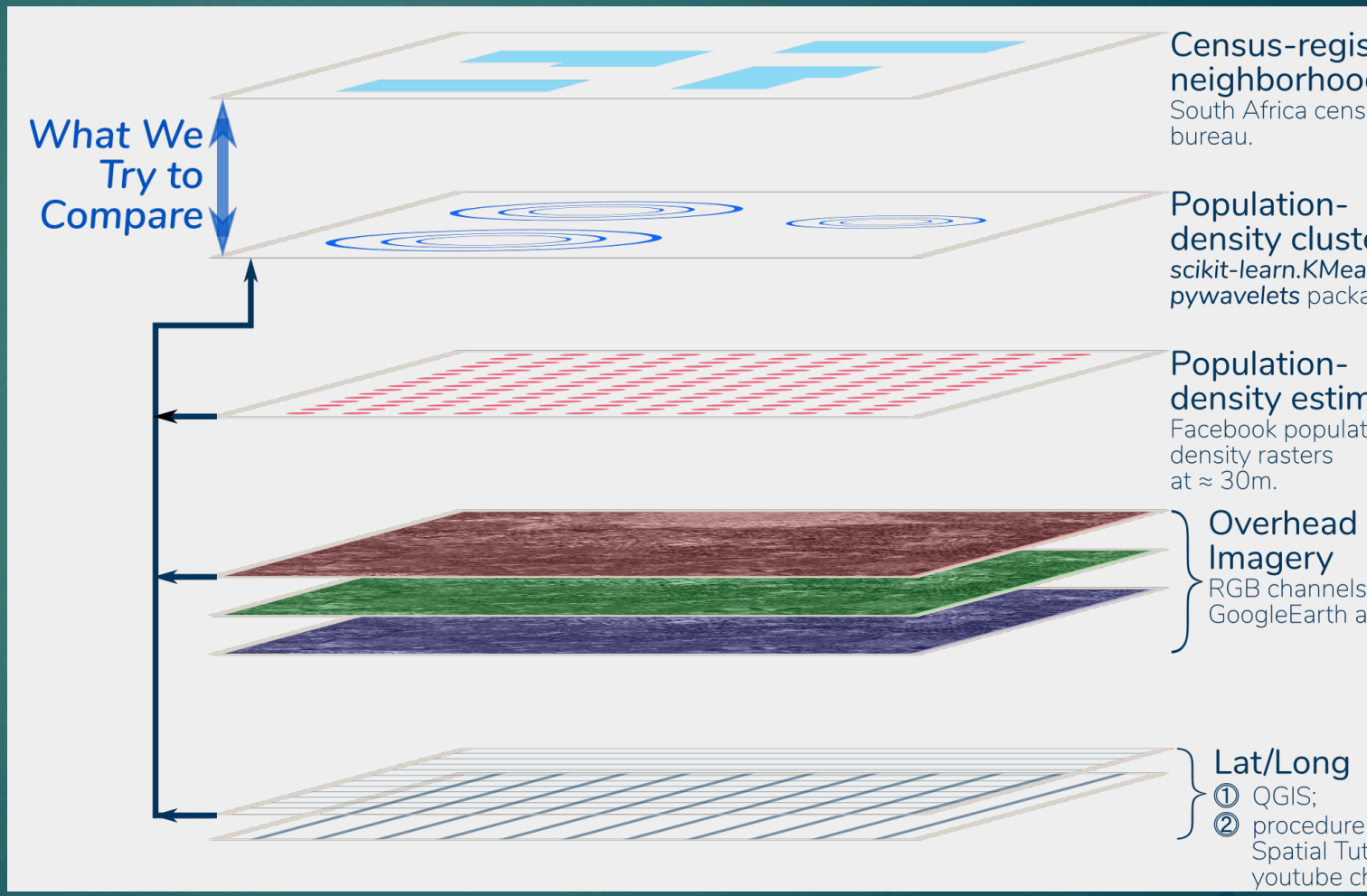
- Partnership with the Center for International Earth Science Information Network (CIESIN) at Columbia University
- Uses neural networks to train pattern recognition algorithms on overheard satellite imagery and then combined with with datasets from the census, USAID, and the World Bank
- Advantages over the census:
  - Updated annually
  - Claims to report population distribution at a 30 meter resolution

s://dataforgood.fb.com/docs/methodology-high-resolution-population-density-maps-demographic-estimates/

echnical Approach

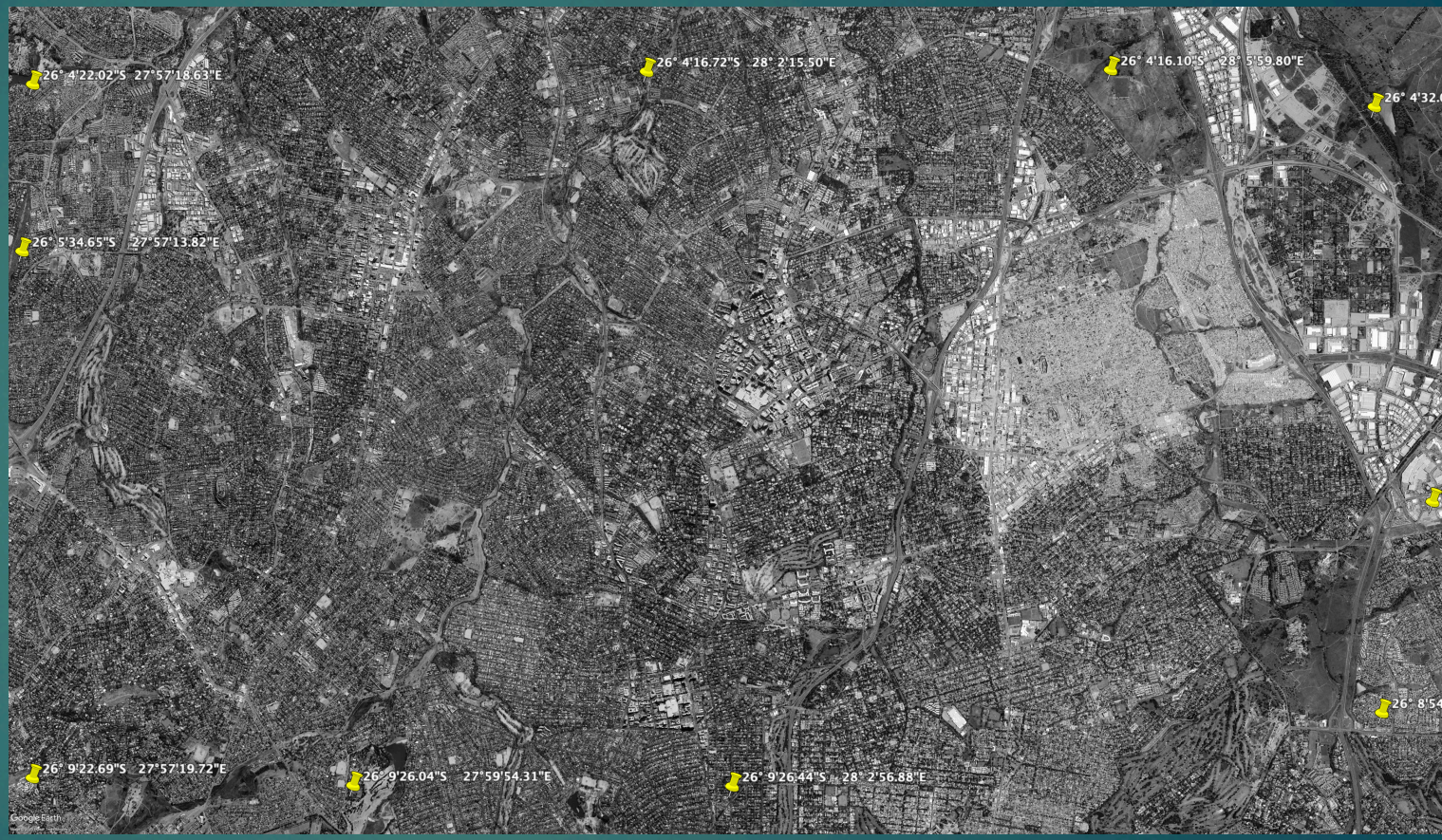
ation Sources:

underlying geopsatial
nce frame;
ficial" population-
ution information; and
ernative" population-
ution information.

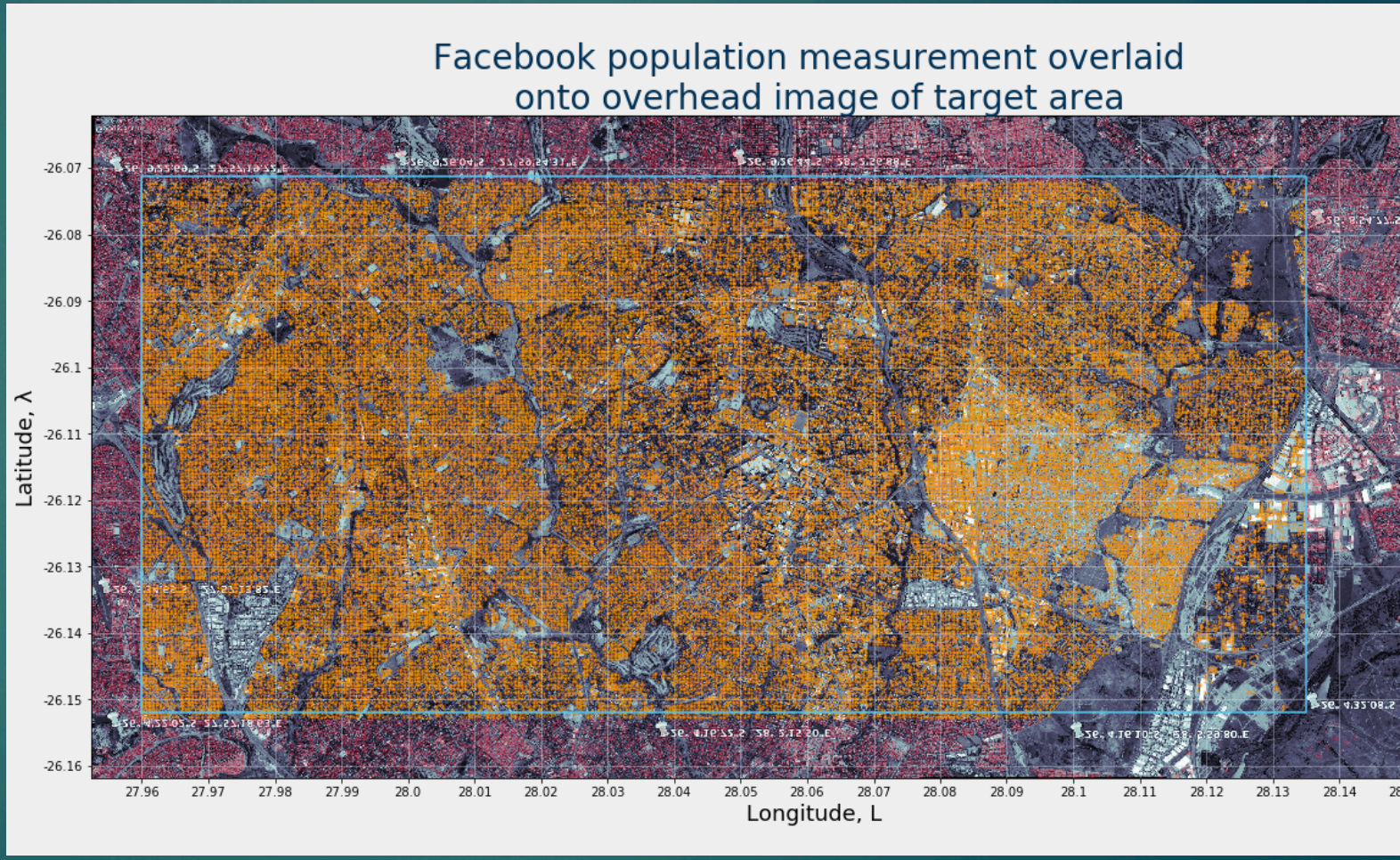


oogle Earth's imagery is
-referenced, we
a limited number of
erenced labels (right)
ed QGIS to attribute (*L*,
ciations to our labels.

oduct: GeoTiff image
estimated pixel
on in the range of five to
ters.



box: windowing,
000 measurements
pixels: estimates in
ns on interest, > 94,000
surements



Modeling the Data

- An unsupervised learning approach gives insight into the features that discriminate between the two as we seek to compare population clusters in our "ambient" data against "official" measurements
- Explanatory variables:
  - Features from our Facebook population-density estimates including both geographic ($L,\ \lambda$) and the actual population estimates;
  - Features extracted from our overhead imagery to introduce spatial cohesion
- Possibly future extension to binary classification.

Modeling the Data: Imagery Feature Extraction

- Discrete Waveform Transformation: recursively partitions a data set a specified number of times. This partition is accomplished using *filters*, which orthogonalize the data at each stage resulting in a set of features that are uncorrelated.
- The distinct Facebook population-density estimates at 30-meter resolution provide our principal explanatory variables. These contain geographic ($L,\ \lambda$) and point population-density estimates. We extend these with imagery-feature attributes. First, we compress our three-channel (Red, Blue Green) image into a single-channel gray-scale intensity array. We take a 128×128-pixel window centered on each Facebook population-density estimate.
- The model described here is based on three levels. This produced an explanatory-variable matrix comprised of 1,603 features for each of ≲66,000 observations.

-Means Clustering

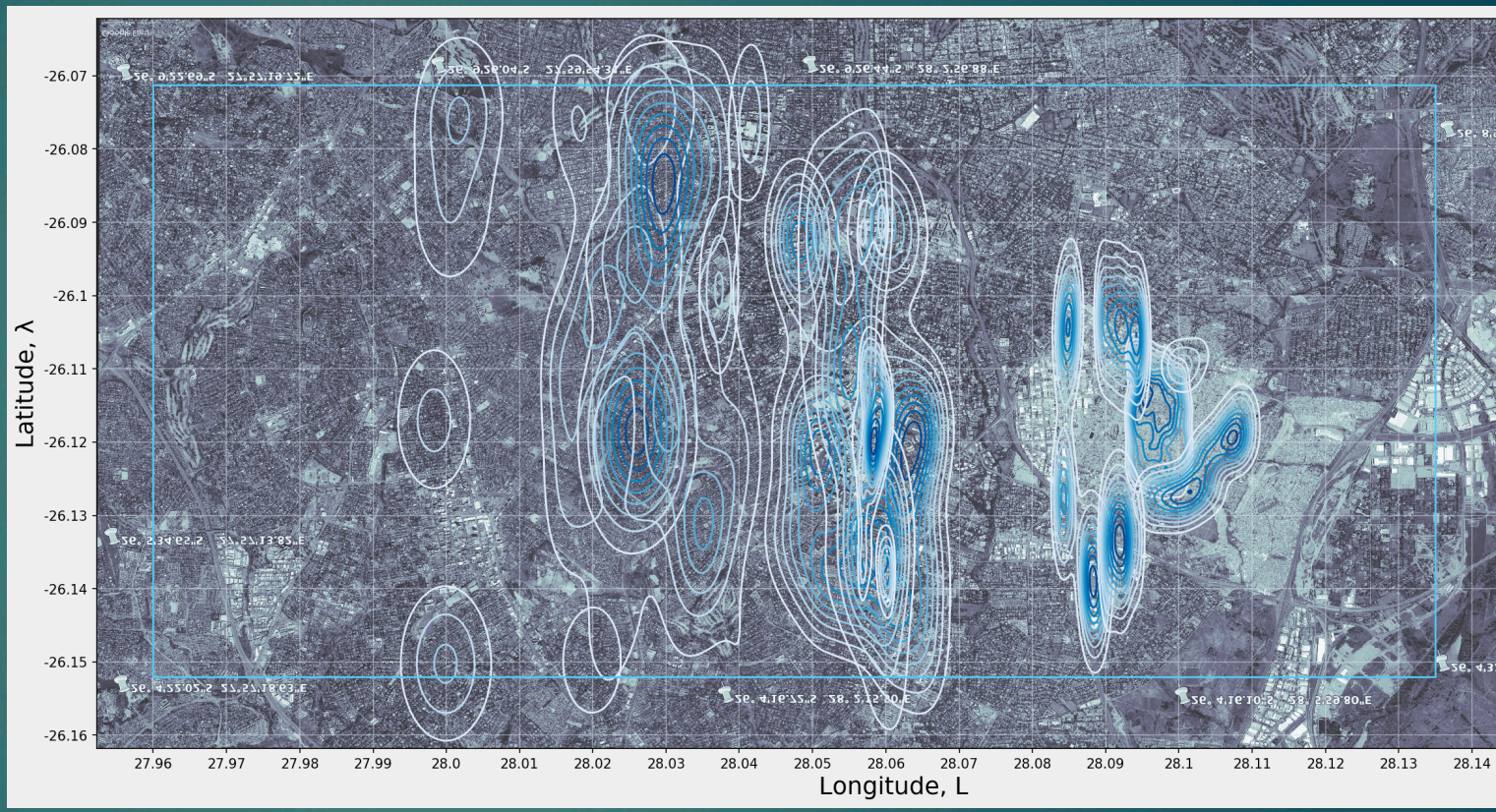
- Explanatory variables included the Facebook population-density estimates, the corresponding geographic ($L,\ \lambda$) for each, and all of the coefficients from a three-level DWT of a 128×128-pixel window centered on each.
- Models were fit for variations on these characteristics, including DWT window size and DWT levels. Models were considered with the data scaled and non-scaled. Non-scaled data produced more-localized results, owing to the affects of strong ($L,\ \lambda$).
- Although we attempted Principal Component Analysis to reduce the number of features, it afforded no reduction in dimensionality.

ernal Density Estimate

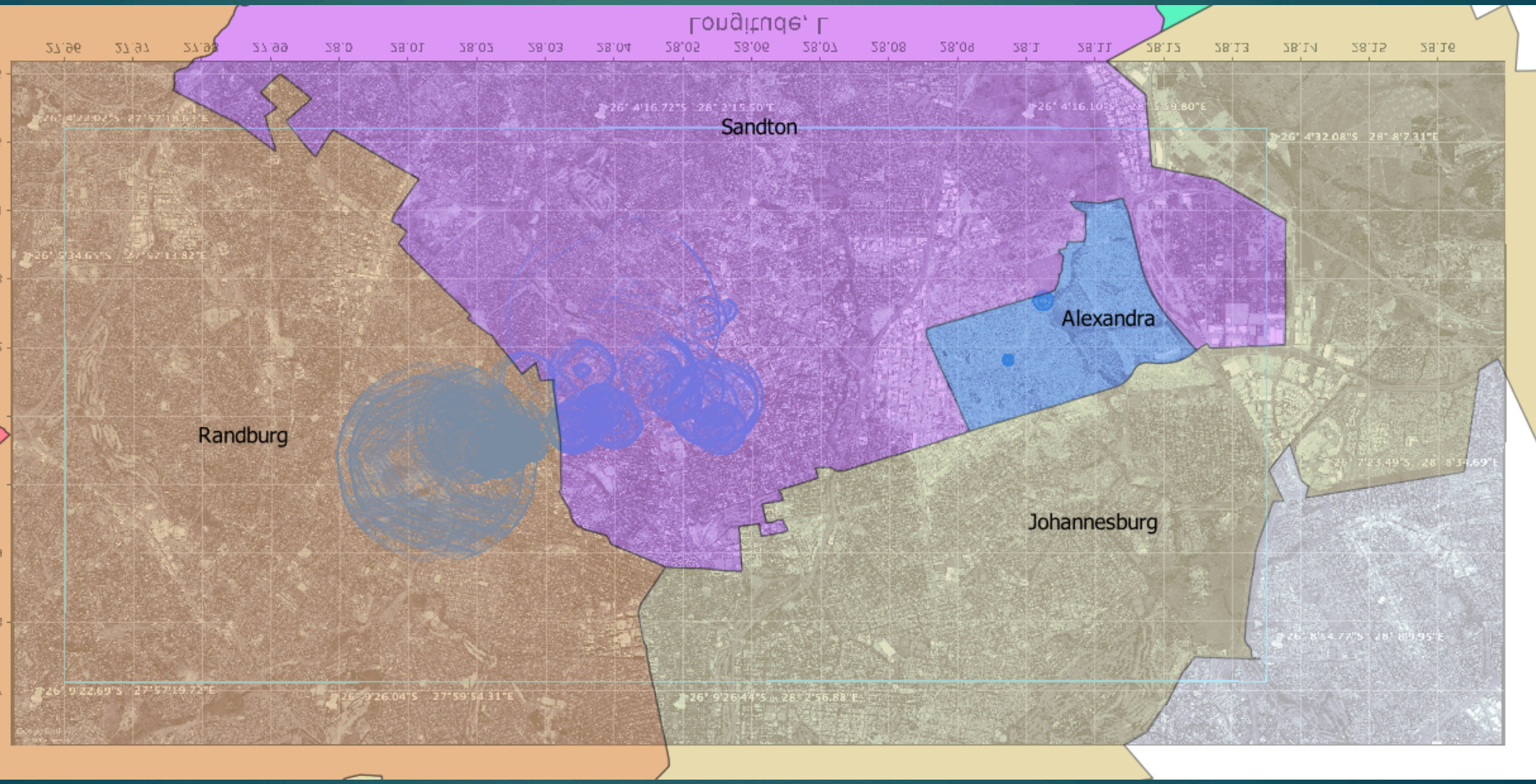
uster distributions are
nvex and widely distributed,
ularly expansive in the vertical
sion with multiple peaks. This
s lots of cluster members are
to be widely separated, with
silhouette scores.

ocalized clusters in the vicinity
know "slum", Alexandra,
ng. This is the bright area
ded by about 28.08 and 28.1 *L*
6.13 and 26.11 *λ* . *Our imagery*
es appear to discriminate slum
on-slum areas.

m contours appear to be
ntrated between about 28.02 and
. This seems to be the heart of the
nesburg populated area. It seems
e Population attribute is dominant
egion.



valuation: Model Contours with Overhead dministrative District Boundaries



Conclusions and Recommendations

- The lack of real estate data makes it more difficult to identify slum areas but using satellite data to locate clusters of upmapped populations in conjunction with official census and open source data is promising.
- Recommendations for Improvement:
  - "Professional grade" satellite imagery vs. Google Earth
  - Harnessing cell phone geolocation to improve identifications of slum areas not thoroughly represented in the census.
  - More-sophisticated machine-learning algorithm.