GIS models for address canvassing - census 2022.

14 June 2023
Topics

1. Census and 2030 vision
2. Address canvassing context
3. The two GIS models
4. Results
Census will be key pillar of Vision 2030 and will generate critical data enabling better decision-making

What is Census 2022?

- KSA's 5th nation-wide census since 1974
- Entire population counted along with key attributes (including housing, education, occupation)
- Effort led by GASTAT, but requires collaboration of all government stakeholders
- Following highest standards of data confidentiality and privacy
- Introducing innovative methods for data collection and dissemination

Why is Census 2022 important?

- Enables Vision 2030 objectives, informs all stakeholders of the direction of the Kingdom’s socioeconomic reforms
- Generates valuable insights into Saudi population and recent transformations or shifts across the Kingdom
- Enables decision-making for important government initiatives (e.g. urban planning, transportation infrastructure, subsidy allocation)
- Paves the way for the future of data collection and dissemination
- Informs private sector on the nature of the Saudi market
Census objectives are improving 2 Vision2030 objectives part of National Transformation Program (NTP) ..

‘Statistical Data & Information Production’ is one of GASTAT strategy dimensions where census program is contributing on achieving governmental operational excellence

**Vision 2030 Objectives**

- An Ambitious Nation
- Achieve Governmental Operational Excellence

**Census Objectives**

**Strategic objectives from GaStat Strategy Dimension:**
Statistical data production (demand)

- Improve the Quality of Services Provided to Citizens
- Enhance Transparency Across Government Entities
- Provide easy to use statistical data, & on time, to meet users requirement
- Enhance quality of statistical products & services

**Census goals:**
Producing census results following internationally agreed census principles

- Enhancing census key dimensions:
  - Address canvassing using geospatial analysis
  - Data collection through digital self-enumeration
  - Increased focus on data quality
  - National media campaign to drive census participation
  - Dissemination with a focus on the end-user
.. and census outcomes enables the rest of the vision 2030 objectives and dimensions

### Census outcomes provide the raw data to:

1. **Enhance inputs for better planning for vision 2030 related studies** across sectors, and
2. **Provide framework for specialized statistical surveys** across sectors (e.g. labor, household spending & income, demographic research, .. etc)

#### Components

**Ambitious Nation**
- Unlocking *government services* to the population, e.g.:
  - Government effectiveness
  - Data for private sector footprint

**Thriving Economy**
- Sizing the *economy* today, e.g.:
  - Attract FDI
  - Enabling SMEs
  - Household income & spending

**Vibrant Society**
- Understanding *social fabric* and *demographical needs*, e.g.:
  - Percentage of citizens owning housing
  - Insights on labor & volunteers

#### Examples

<table>
<thead>
<tr>
<th>Components</th>
<th>Examples</th>
</tr>
</thead>
</table>
| **Ambitious Nation** | Unlocking *government services* to the population, e.g.:
|  |  • Government effectiveness
|  |  • Data for private sector footprint |
| **Thriving Economy** | Sizing the *economy* today, e.g.:
|  |  • Attract FDI
|  |  • Enabling SMEs
|  |  • Household income & spending |
| **Vibrant Society** | Understanding *social fabric* and *demographical needs*, e.g.:
|  |  • Percentage of citizens owning housing
|  |  • Insights on labor & volunteers |

**Regional insights**
1. A comparison between the traditional approach and the hybrid approach in identifying addresses

### Traditional approach
**Field workers only**
- The implementation approach is widely known by the Ops team

### Hybrid approach
**Field workers + GIS**
- Robust approach to find new addresses, specially in rural areas
- GIS model can be used to find new addresses in the years post-census

### Pros
- More prone to have **missing addresses**
- Requires high technical experience and time to make the first iterations of the models
- Needs coordination between GIS and Ops teams

### Example of missing addresses
2. The GIS model based on open datasets make use of more than 10 data sources

List of 10 of the external datasets

<table>
<thead>
<tr>
<th>External data</th>
<th>Source</th>
<th>Recency (date)</th>
<th>Granularity</th>
<th>Description of data</th>
<th>How data is collected?</th>
<th>How we propose to use the data?</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building footprints and population</td>
<td>HRSL</td>
<td>2020</td>
<td>30m</td>
<td>Building footprints from satellite imagery (dataset contains only pixels where buildings were detected)</td>
<td>Satellite imagery analysis using deep learning</td>
<td>To find missed addresses in the old database (especially for rural areas), and to estimate population in missed or new addresses whenever GASTAT or electricity data is missing</td>
<td>~100%</td>
</tr>
<tr>
<td>Building footprints and population</td>
<td>WorldPop</td>
<td>2020</td>
<td>100m</td>
<td>Population estimation in areas containing buildings (dataset contains only pixels where buildings were detected)</td>
<td>Past KSA census + machine learning algorithms + UN pop estimates</td>
<td>To estimate population in missed or new addresses whenever GASTAT or electricity data is missing</td>
<td>~100%</td>
</tr>
<tr>
<td>Building footprints</td>
<td>GHS</td>
<td>2018</td>
<td>10m</td>
<td>Building footprints from satellite imagery (dataset contains only pixels where buildings were detected)</td>
<td>European Commission satellite imagery analysis + machine learning</td>
<td>To find missed addresses in old database (especially for rural areas)</td>
<td>~100%</td>
</tr>
<tr>
<td>Population</td>
<td>GHS</td>
<td>2019</td>
<td>250m</td>
<td>Population geospatial grid</td>
<td>European Commission satellite imagery analysis + machine learning</td>
<td>To estimate population in missed or new addresses whenever GASTAT or electricity data is missing</td>
<td>~100%</td>
</tr>
<tr>
<td>Urbanicity</td>
<td>GHS</td>
<td>2015</td>
<td>1km</td>
<td>Geospatial grid containing urbanicity level (from very rural to very urban)</td>
<td>European Commission satellite imagery analysis + machine learning</td>
<td>To separate analyses between urban and rural areas</td>
<td>~100%</td>
</tr>
<tr>
<td>Building footprints</td>
<td>OSM</td>
<td>2021</td>
<td>Actual location</td>
<td>Building footprints and residential areas as polygons</td>
<td>Open source system + volunteers</td>
<td>To find missed addresses in old database or find new addresses that were created in last 11 months (especially for rural areas)</td>
<td>~20%</td>
</tr>
<tr>
<td>Roads</td>
<td>OSM</td>
<td>2021</td>
<td>Actual location</td>
<td>All roads in KSA</td>
<td>Open source system + volunteers</td>
<td>To separate analyses between close and far away from roads + calculating road distances between addresses and closest cities</td>
<td>~100%</td>
</tr>
<tr>
<td>Nightlights</td>
<td>VIIRS-NOAA/NASA</td>
<td>2016</td>
<td>500m</td>
<td>Nightlights from satellite imagery (each pixel has a light intensity value from 0 to 255)</td>
<td>Satellite imagery analysis of photos throughout the whole year</td>
<td>To find missed addresses in old database (especially for rural areas)</td>
<td>~100%</td>
</tr>
<tr>
<td>Cities locations</td>
<td>HDX</td>
<td>2020</td>
<td>Actual location</td>
<td>Coordinate of cities</td>
<td></td>
<td>To separate analyses between urban and rural areas + calculating road distances between addresses and closest cities</td>
<td></td>
</tr>
<tr>
<td>Cell towers</td>
<td>OpenCellID</td>
<td>2021</td>
<td>Actual location</td>
<td>Location of cell towers</td>
<td>Volunteers, phone apps and private contributors</td>
<td>To find missed addresses in old database (especially for rural areas)</td>
<td></td>
</tr>
</tbody>
</table>
2. GIS model based on open datasets: Rural and urban models were developed using the open datasets to find new addresses

**Technical details**

**Approach:** decision tree ensemble model for estimating probability of building in each 100mx100m grid cell

**Model details:**
- Hyperparameter tuning
- Cross validation
- Threshold tuning by geographical area using precision-recall curves
- Probability calibration test

**Coding:** Python, Scikit-learn

**Data sources:** 10 open datasets mapped to 100mx100m grid

**Data size:** model applied to 200 million grid cells of 100mx100m

**Example heatmap of building probabilities**

- Probability of building
  - 0 - 0.1
  - 0.1 - 0.2
  - 0.2 - 0.3
  - 0.3 - 0.4
  - 0.4 - 0.5
  - 0.5 - 0.6
  - 0.6 - 0.7
  - 0.7 - 0.8
  - 0.8 - 0.9
  - 0.9 - 1
2. GIS model based on satellite imagery: A deep learning model was developed to find buildings in high resolution satellite imagery

**Technical details**

**Approach:** deep learning CNN for estimating probability of building in each 100mx100m grid cell

**Model details:**
- ResNet50
- Binary cross-entropy

**Coding:** Python, Keras

**Data sources:** high resolution satellite images for 5 cities

**Data size:** model applied to classify ~2 million image chips of 100mx100m
3. potential new addresses\(^1\) were derived from the GIS models after applying two address filters

**Filtering sequence**

<table>
<thead>
<tr>
<th>Addresses generated by the two GIS models not overlapping GASTAT ones</th>
<th>Filter 1</th>
<th>Filter 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keep only addresses far away from existing addresses(^1)</td>
<td>Human desktop review</td>
<td></td>
</tr>
</tbody>
</table>

GIS models outcome

The GIS models potential new addresses are delivered to the GASTAT team for review during address canvassing

---

1. >100m away from existing GASTAT addresses in urban areas and >200m away in rural areas
3. A human desktop review was performed to increase the quality of the potential new addresses

Why a human review is needed?

- To move pins to exact building location
- To eliminate agricultural structures pins
- To add more nearby buildings
- To merge pins into single pin
Thank You