Arab Thematic Conference on Agile and Resilient
National Statistical Systems - Amman, Jordan

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

الهيئة العامة للإحصاء
General Authority for Statistics

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) ..

الهيئة العامة للإحصاء
General Authority for Statistics

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

Vision 2030 Objectives

Theme	L2	L3
An Ambitious Nation	Achieve Governmental Operational Excellence	Improve the Quality of Services Provided to Citizens
		Enhance Transparency Across Government Entities

Census Objectives

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

- 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

الهيئة العامة للإحصاء
General Authority for Statistics

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)

Examples

Components

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

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

- 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

الهيئة العامة للإحصاء
General Authority for Statistics

	Pros	Cons
Traditional approach **Field workers only**	The implementation approach is widely known by the Ops team	More prone to have **missing addresses**
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	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

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

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

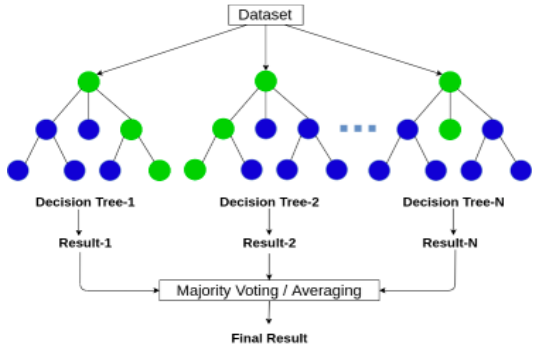
Model details:

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

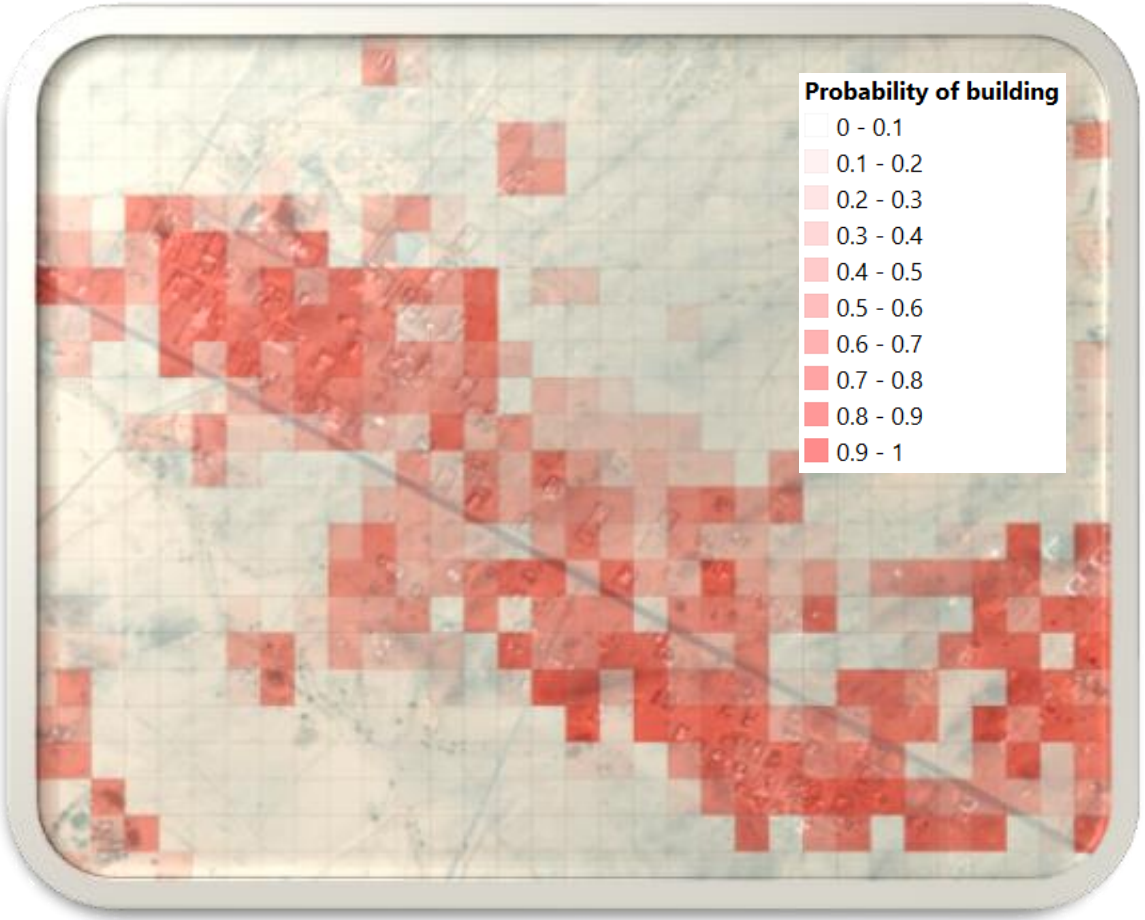
Coding: Python, Scikit-learn

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



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

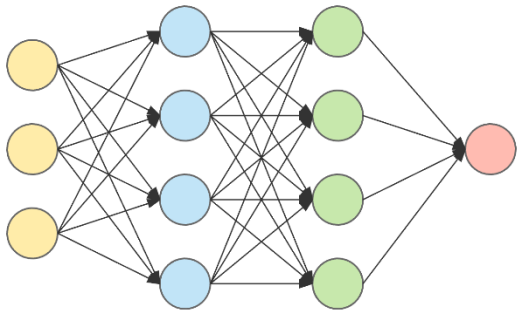
Model details:

- ResNet50
- Binary cross-entropy

Coding: Python, Keras

Data sources: high resolution satellite images for 5 cities

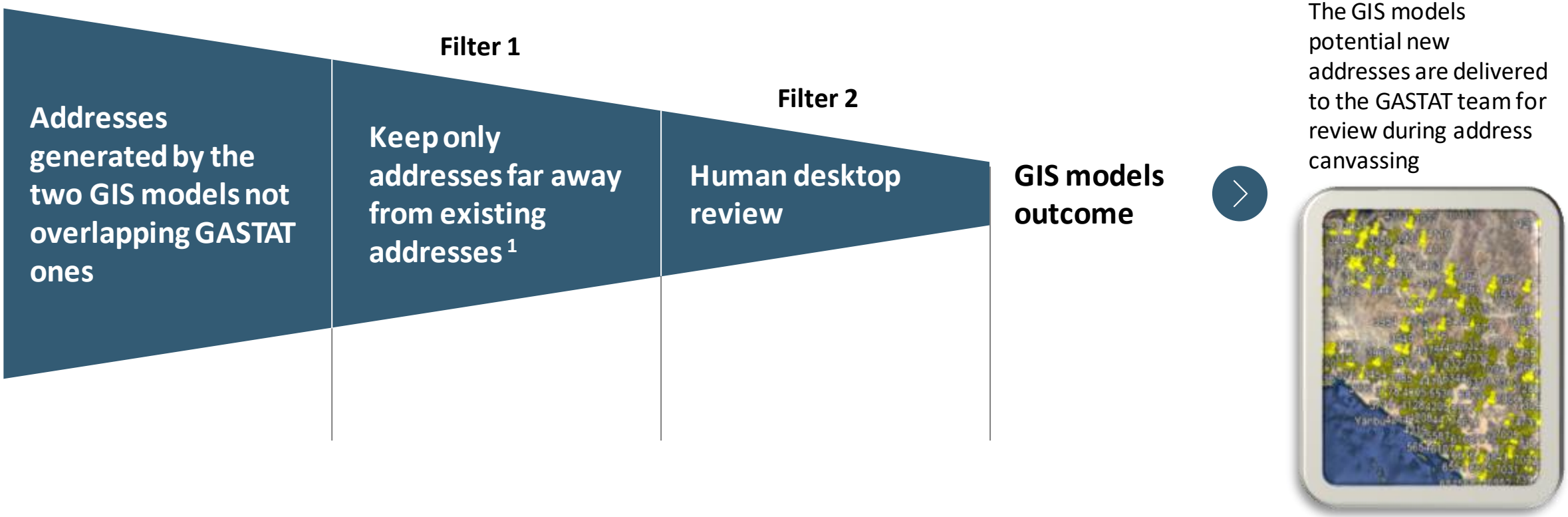
Data size: model applied to classify ~2 million image chips of 100mx100m



Example of buildings detected

3. potential new addresses[1] were derived from the GIS models after applying two address filters

Filtering sequence



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

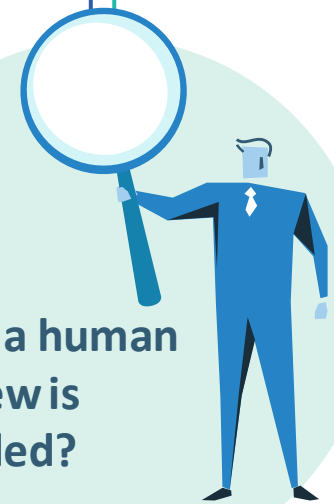
Pin from model Pin after human review

الهيئة العامة للإحصاء
General Authority for Statistics

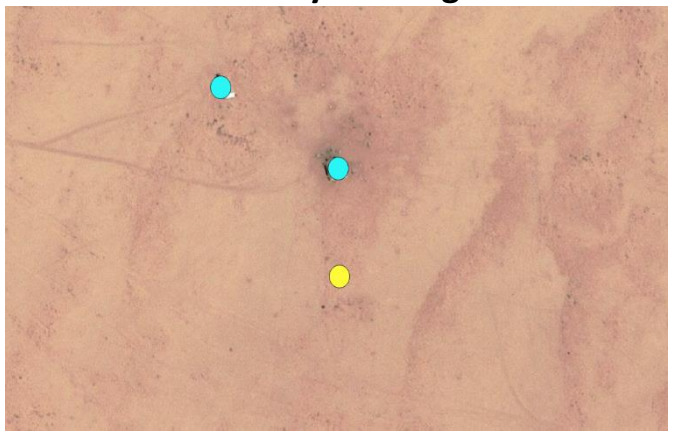
To move pins to exact building location

To eliminate agricultural structures pins

Why a human review is needed?



To add more nearby buildings



To merge pins into single pin

Thank You