Mining Satellite Images for Census Data Collection: A Study Using the Google Static Maps Service

Frans Coenen

(http://cgi.csc.liv.ac.uk/~frans/)

UNIVERSITY OF LIVERPOOL

IC3K 2016

Porto, Portugal, 9-11 November 2016

Acknowledgements

Kwankamon Dittakan

Faculty of Technology and Environment

Prince of Songkla University (PSU)

Thailand

Census Data

- A census is a mechanism for acquiring and collecting information about a population.
- Widely used with respect to a variety of national. and local government, management and planning activities.
- Most important element of a census is population count.

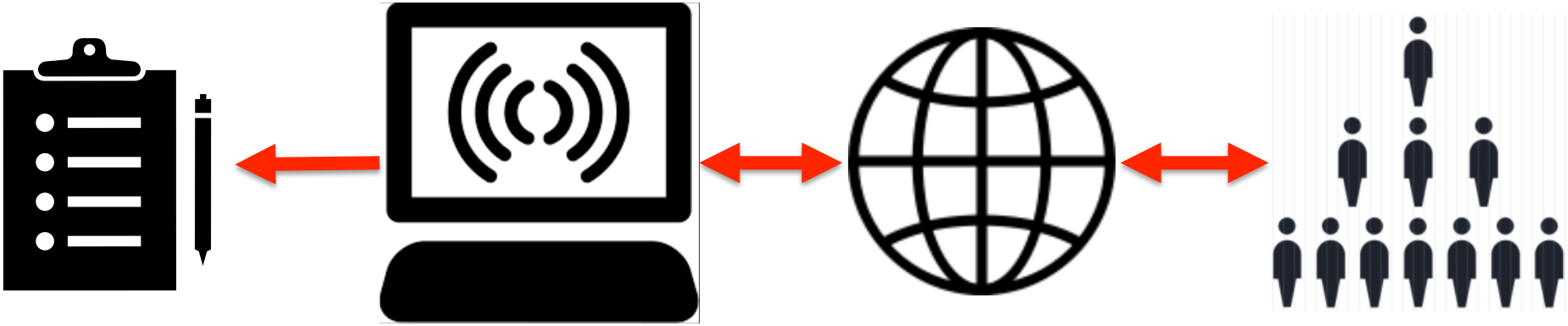
Liverpool
Population = 465,700
(2011 census)

Challenges of Census Collection

- Census collection and post processing of data is expensive:
  - The UK Office for National Statistics (UKONS) reports that the UK 2011 census cost some £480 million.
  - The US 2010 census is reported to have cost $13 billion.
- The cost of census collection is increasing:
  - According to the Australian Bureau of Statistics the Australian 2006 census cost around AUD 300 million; whilst the 2011 census cost around AUD 440 million.
- Cost with respect to rural areas is typically greater than in urban areas because the communication and transport infrastructure in rural areas tends to be less well developed.
- There is also often a lack of good will on behalf of a population to participate in a census, even if they are legally required to do so, because people are often suspicious of the motivation behind censuses.

Solution One: Technology Utilisation?

- Usage of technology, namely the internet. However:
  - Many people remain unconnected to the internet. In the context of the UK 2011 census it was found that the most frequently cited reason for households not to have internet access was because of a “life style” decision not to.
  - In less affluent parts of the world internet accessibility and usage is much lower (although arguably set to increase).
  - Internet based census collection requires those completing the questionnaires to be literate, not necessarily always the case.



Solution Two: Areal Interpolation for Population Estimation?

- Population estimation has been a subject of researched amongst the Geographic Information Systems (GIS) and remote sensing communities for some time.
- In areal interpolation existing census information concerning some geographic area is used as an input to an interpolation algorithm to obtain a population estimation for a wider or alternative geographic area.
- We define an area for which we know the population size according to some set of relevant attributes and then (say) perform linear regression to produce a model that we can then used with respect to other areas that subscribe to the same attribute set.
- This seem like a good idea, however, the question remains as to what this attribute set should be comprised off, we need an attribute set based on information that is readily available at low cost.

Solution Three: Remote Sensor Based Population Estimation

- Use statistical modelling to determine the relationship between population size/density and data obtained from GIS and/or satellite imagery.

- Statistical models can be built using: (i) light intensity, (ii) land usage, (iii) dwelling unit count and (iv) image pixel characteristics.

- Typically done at a region/area level of granularity

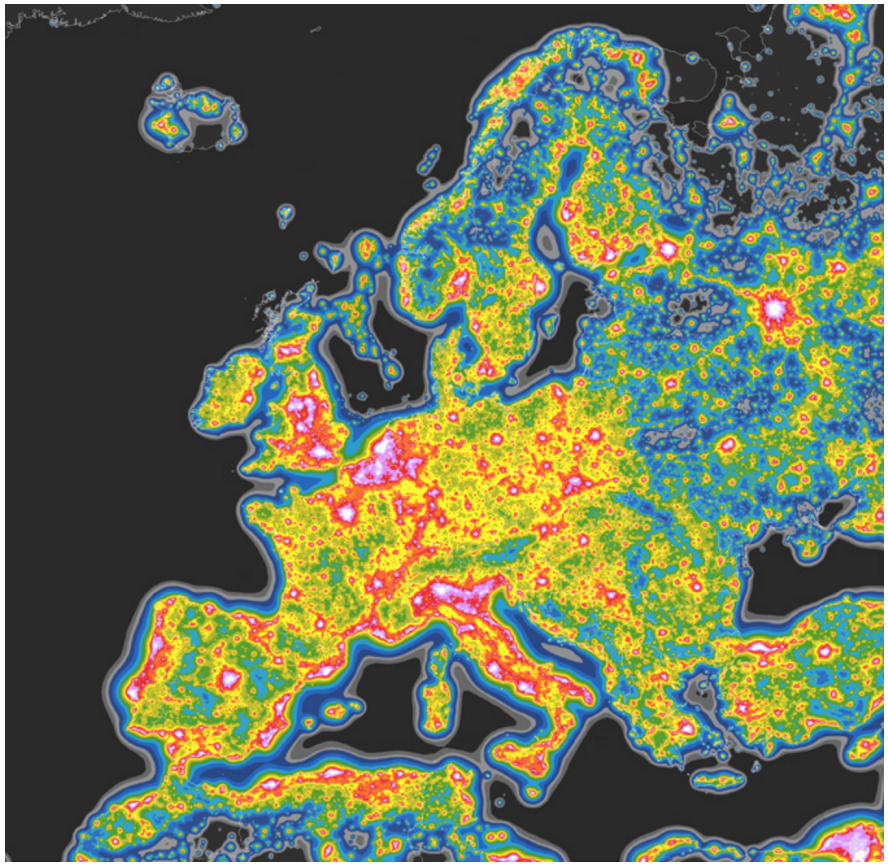


Image from: http://d3a5ak6v9sb99l.cloudfront.net/content/advances/2/6/e1600377/F5.large.jpg

Proposed Solution

- Use satellite image data from which house holds can be isolated (segmented).
- Represent households in some manner to generate feature vectors that can be related to known household size.
- Build a classification model that can be used to predict household size.

Disadvantages:

- Not going to work in cities where difficult to distinguish buildings in terms of number of inhabitants, but will work well in rural areas where census data collection tends to be more of a challenge.
- Need training data. 

Application domain

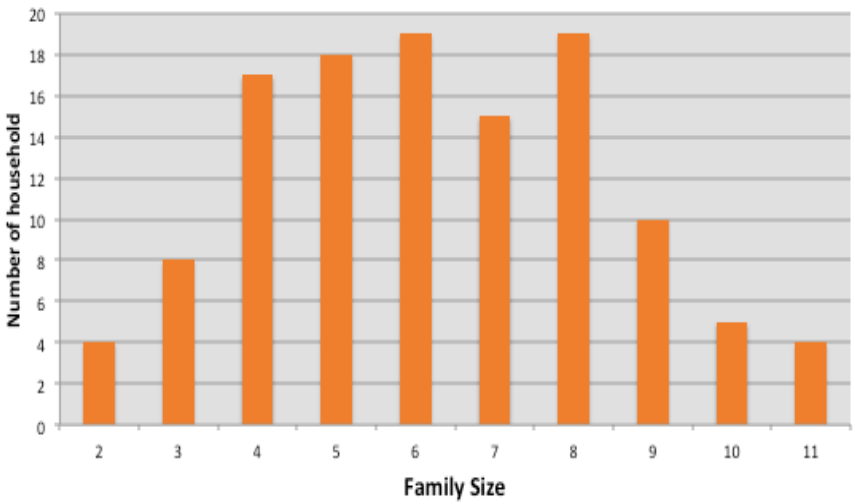
- Horro, Ethiopia.

Application domain cont.

Data Collection

Data Collection cont.

Family Size	Min	Max	Ave	Mode	Site A (wet)	Site B (dry)
Small	2	5	4.04	5	38	19
Medium	6	8	7.00	6	32	21
Large	9	12	9.80	9	10	10
Total 120	2	12	6.31	6	70	50



Training Data Generation

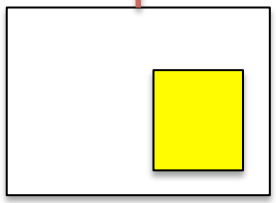
- Locations recorded for households where household size was known, thus can obtain relevant Google Earth satellite images.
- Google Earth does not readily facilitate the automated extraction of satellite imagery, instead used the Google Static Map Service.
- This features an API that allows users to download satellite images (one image at a time) specified according to various parameter settings.
    1. Latitude and longitude of centre of area of interest.
    2. Image size (in pixels).
    3. The Zoom Level (level of detail).
- We used image size of *1280×1280* pixels (*k×LevelOfDecomp*) and zoom level of *18*.
- Surrounded each household with a 256*×256 pixel bounding box defined so as to cover average household (*By superimposing a box we do not have issues with irregular shaped household plots).
- In this manner produced a set of household images

Image Representation for KDD

Image Rep. for KD

Global (Whole Image based)

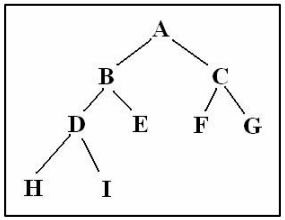
Local (Region/Object based)



Statistical

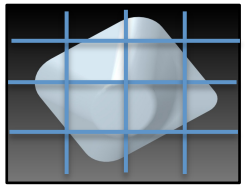
$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Graph based



Individual (regions/ objects)

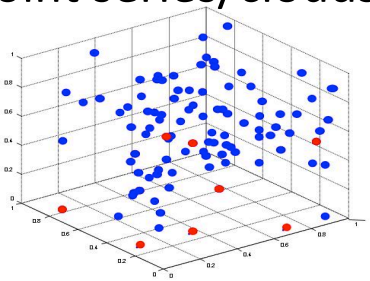
Set of regions objects



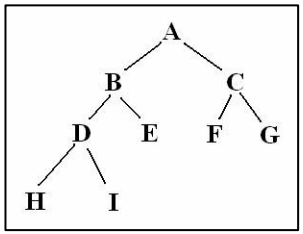
Statistical

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Point series/clouds



Graph based



Statistical

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Image Representation for KDD cont.

- Two approaches: (i) Global (whole image based) and (ii) Local (region or object based).
- The latter, as in the case of the population estimation application, require segmentation.
- Three (broad) techniques:

1. Statistical.
2. Histogram (point series/curves).
3. Graph.

Statistical Techniques

- Simplest approach and easy to represent, in terms of a feature space, directly compatible with classifier generation/application.
- Applied globally or locally.
  - ✓ Morphometrics for local representation.
  - ✓ First Order Statistical functions such as the mean, variance and standard deviation of the intensity or RGB colour values.
  - ✓ Second order statistical functions applied to an intermediate representation (co-occurrence matrices, gradient analysis, Hough transforms).
- General applicability (good benchmark for experimental work).

Histograms

- Many second order statistical techniques lend themselves to representation in the form of histograms
- For example histogram of intensity values, Local Binary Patterns (LBPs) or orientation gradients.
- Histograms can of course be directly translated into a feature vector representation.
- Alternatively, they can be viewed as point series or point curves.

LBPs

p_1	p_2	p_3
p_8	p_c	p_4
p_7	p_6	p_5

Compare each P_c to neighbouring P_i to get a binary digit d:

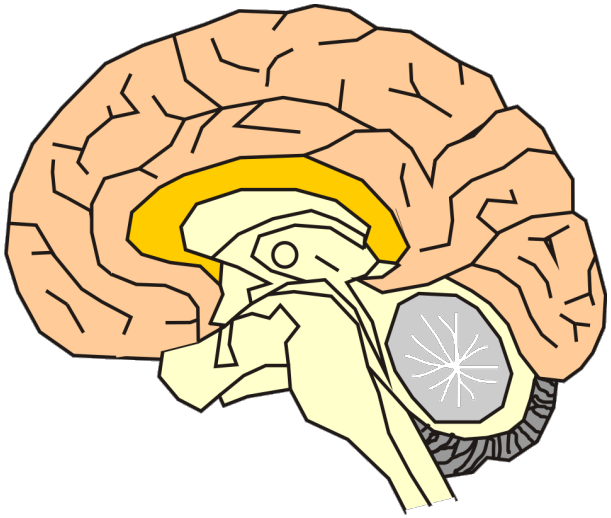
$$d = \begin{cases} 1 & if \quad p_i \leq p_c \\ 0 & if \quad p_i > p_c \end{cases}$$

01101011

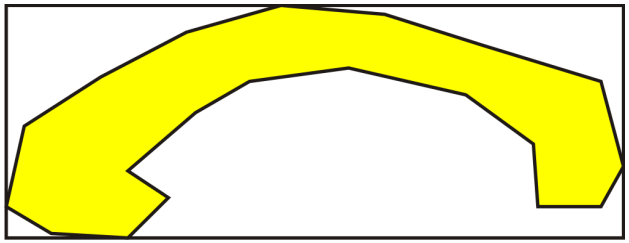
Tree and Graph Based Techniques

- A popular method for representing images is to apply some form of hierarchical decomposition and to store the result in a quad-tree (2D image data) or oct-tree (3D image data).
- Issues with:
    1. "boundary problem" where objects appear in different branches of the tree.
    2. When to stop the decomposition (critical function to measure homogeneity or a pre-specified maximum level of decomposition).

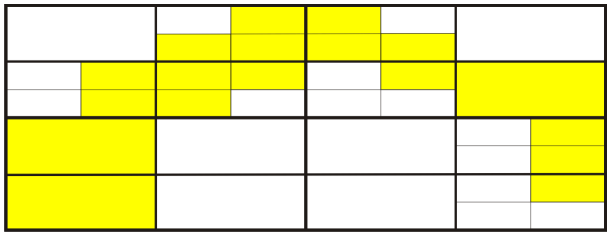
Example Decomposition One



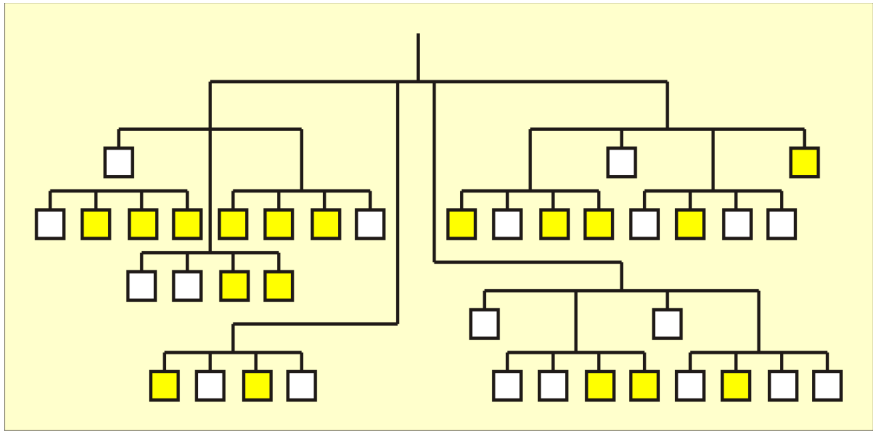
(a) Corpus callosum in a 2D MRI brain scan.



(b) Segmented Corpus callosum.

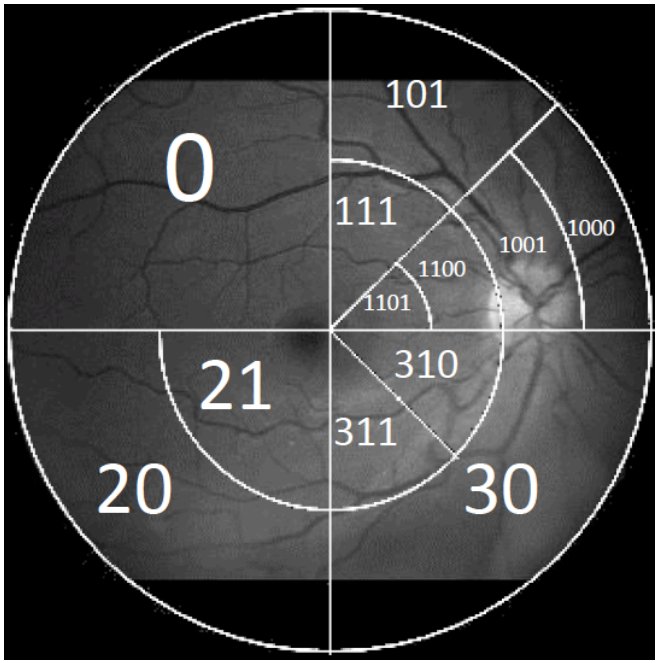


(c) Decomposed Corpus Callosum (level = 3).

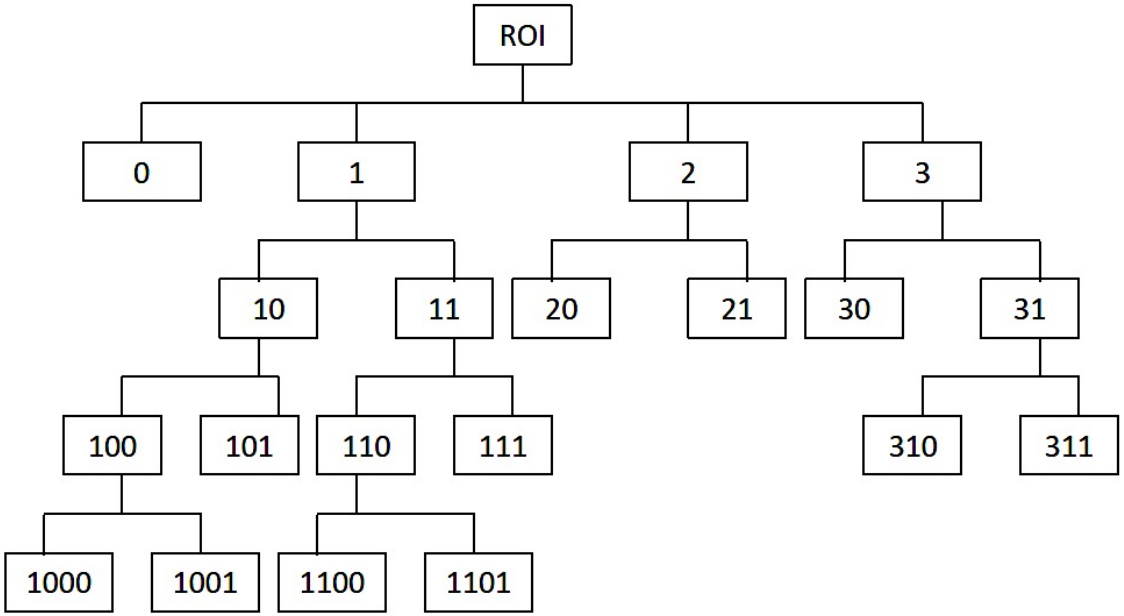


(d) Quad tree representation of the Corpus Callosum.

Example Decomposition Two



(a) Whole image decomposition using an alternative approach (level=4).

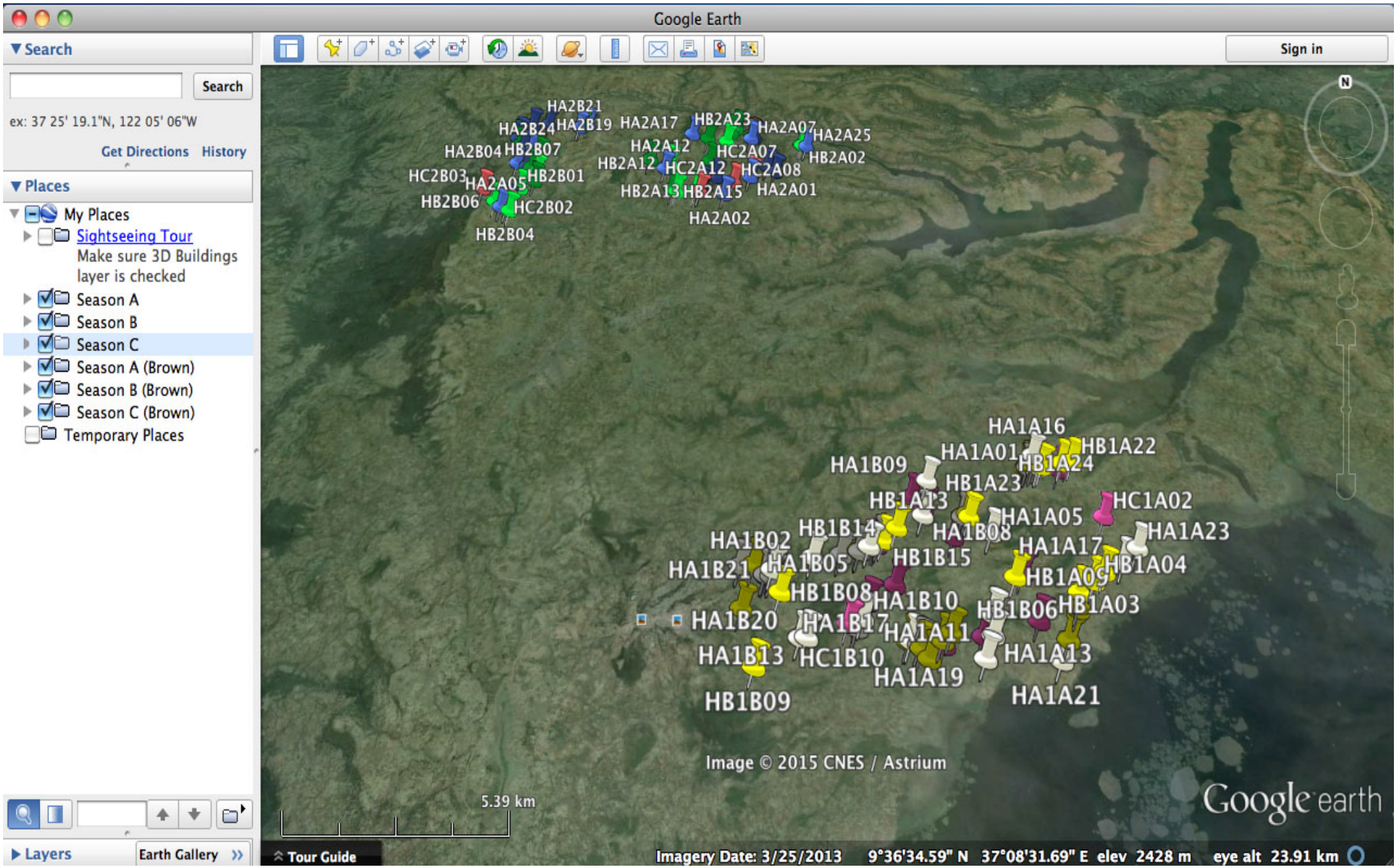


(b) Resulting tree representation.

Generating a Feature Space from a Collection of Graphs

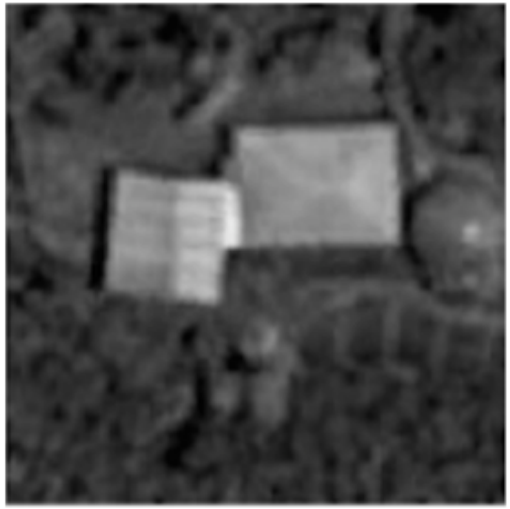
- Apply a Frequent Sub-graph Mining (FSM) algorithm to the data. Frequent defined by some threshold σ. Popular FSM algorithm is gSpan.
- Each frequently occurring sub graph is then a dimension in a feature space.
- This feature space can then be used define feature vectors (binary or real valued) for the initial image set, which can be input to any number of classifier generators (feature selection may also be applied).

Back to Population Estimation

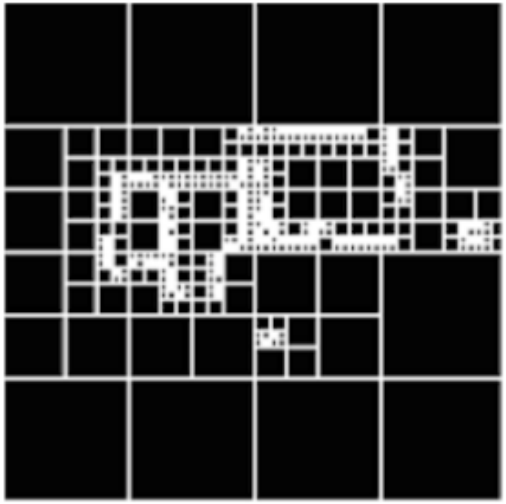


Graph Based

- Generate a set of quad trees (one per household image).
- Apply Frequent Sub-Graph mining to the tree using a support threshold σ (we used a variation of gSpan, low σ values are better, but more FSGs).
- Use frequent sub-graphs to generate feature vectors.
- Apply feature selection.
- Apply your favourite classifier model generator.



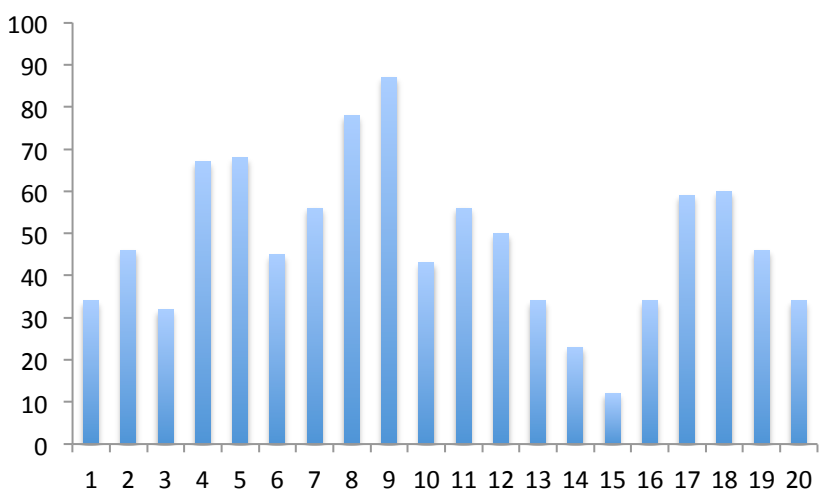
(a) Example of satellite image for a household



(b) Example of satellite image after quadtree decomposition

Colour Histograms

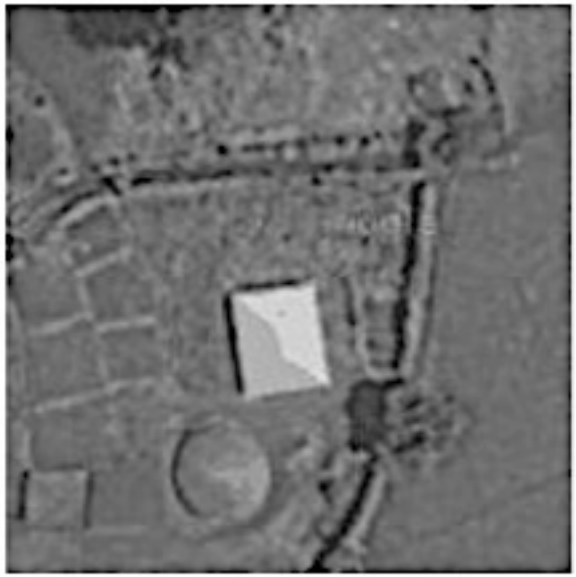
- Generate seven different histograms (red green blue, hue, saturation, value, grayscale), 32 bins per histogram and concatenate them together.
- 224 attribute feature vectors (*32×7=224*), one per household.
- Experimented with including statistical measures but made no difference (detrimental in some areas).
- Considered a number of feature selection metods (χ^2, Gain ratio, Information gain).



Texture Based (LBPs) histograms

- Used LBPs with eight neighbours and a radius of 1.
- Again applied feature selection.
- Experimented with other LBP configurations but no advantage.
- Experimented with idea of including statistical texture metrics as well (contrast, correlation, energy, homogeneity) but little effect.

(a) Segmented household



(b) Grayscale image

(c) LBP image

Evaluation (TCV)

Classification Model Generator	Site A (Wet Season)					Site B (Dry Season)				
	AC	AUC	FM	SN	SP	AC	AUC	FM	SN	SP
Graph based (BN)	0.600	0.808	0.596	0.600	0.734	**0.800**	**0.879**	**0.792**	**0.800**	**0.876**
Graph based (NN)	0.686	0.819	0.685	0.686	0.782	0.620	0.789	0.628	0.620	0.829
H'gram based (BN)	0.700	0.807	0.687	0.700	0.782	0.700	0.798	0.692	0.700	0.829
H'gram based (LR)	0.657	0.822	0.662	0.657	0.806	0.640	0.821	0.633	0.640	0.798
Texture based (LR)	**0.771**	0.859	**0.778**	**0.771**	**0.885**	0.680	0.756	0.679	0.680	0.803
Texture based (NN)	**0.771**	**0.881**	0.759	**0.771**	0.852	0.720	0.824	0.718	0.720	0.825

- Graph Based: σ=10 for FSG mining, Gain ration Feature Selection with k=55.
- Histogram Based: Gain ration Feature Selection with k=25.
- Texture Based: χ^2 feature selection with k=40.

BN = Bayesian Network model, NN = Neural Network, LR = Logistic Regression.

AC = Accuracy. AUC = Area Under receiver operating Characteristic, SN = Sensitivity, SP = Specificity., FM = F-measure (FM).

Large Scale Study

- Once model has been built it can be applied to a much wider area.

Test Area

- Test area chosen because:
  a) Similar area from which the **Site A** (wet) and **Site B** (dry) data was extracted from.
  b) Thus models generated using the Site A and Site B data sets can be used.
  c) Featured a village, and its surrounding lands, that in 2011 was reported to comprise 459 households and a population of 3,223 (thus "ground truth" data was available).
- 600 Satellite images were collected covering the area using the Google Static Map Service API (took 356 seconds).
- Satellite mage data from 2013, two years after the census!

Map Collection

- Used the Google Static Map Service API; image size of *1280×1280* pixels and zoom level = *18*.
- Images downloaded in an iterative manner, image by image, using a 320 pixel overlap (overlap designed so that every household will appear in its entirety in at least one image).
- For this to operate correctly it was necessary to: (i) convert the top-left corner lat. and long. of the current image into *x* and *y* pixel values, (ii) add the required offset to obtain the top-left *x* and *y* coordinates of the next image in the sequence, (iii) convert these new *x* and *y* coordinates back to a latitude and longitude and (iv) repeat.
- Cartesian coordinates are planer values while lat. and long. are geoidal, so conversion not straight forward.
- Google Static Map Service uses EGM96 (Earth Gravitational Model 1996).

Map Collection cont.

Image Segmentation

- Downloaded satellite images could contain zero, one or more households.
- Segmentation conducted using a number of image masks.
- Experiments conducted using a variety of image formats and masking techniques (a significant challenge was the illumination of roads and water ways).
- Found that masks expressed in terms of the HSV (Hue-Saturation-Value) colour space produced the best results

Image Segmentation cont.

(a) Original image (RGB)

(b) HSV colour space image

(c) Hue channel mask

(d) Saturation channel mask

(e) Value channel mask

(f) Intersection HSV mask

Household Data Set

- After segmentation we have a set of households images each identified by a central latitude and longitude surrounded by a w×*w* box (w=256, same value as used for classification model training).

- Boxes will be smaller and/or non-symmetrical near edge of each image.

- Use knowledge of Latitude and Longitude, and box size, to remove duplicate household images.

Household Detection

- 526 households detected including duplicates (processing time 1,370 seconds (22.8 minutes) about 2.28 seconds per satellite image and 2.6 seconds per household.
- Duplicate detection identified 100 duplicate households, thus 426 out of a "known" number of 459 households were identified.
- Suggested reasons for the discrepancy were as follows:
  1. Two year time difference between "ground truth" survey and satellite images; a period during which some households may have fallen into disuse (manual inspection of a proportion of the collected satellite images indicated that some households did indeed appear to be roofless thus supporting this conjecture).
  2. Inspect of the satellite imagery indicated that a small number of buildings were very poorly defined and in some cases not segmented correctly.
  3. It was also possible that the duplicate household detection mechanism had detected some duplicates that were in fact not duplicates (although no evidence for this was found).

Results

Prediction Model	Population Estimation	Accuracy (%)	Total Run Time (Minutes)
Neural Networks classification with Chi-Squared feature selection and LBP (Site A wet season data).	2,545	78.96	29.49
Bayesian Network classification with Gain Ratio feature selection and graph-based representation (Site B dry season data).	2,495	77.41	35.42
SVMreg with CFS feature selection and LBP representation (Site A wet season data).	2,548	79.06	29.48
SVMreg with CFS feature selection and LBP representation (Site B dry season data).	**2,760**	**85.63**	29.48

Discussion

- The data from which the prediction models were generated might not reflect the data to which they were applied as closely as anticipated. Measures for determining the similarity between satellite image data sets are a subject for future work.

- Two year time gap between the date of the census collection (2011) and the date of the satellite image extraction (2013). Manual inspection of a number of images showed signs of derelict (abandoned) households. It may thus be the case that between 2011 and 2013 depopulation had taken place and that the produced population estimates were in fact a better reflection of population size than initially thought (recent reports on depopulation in rural Ethiopia).

- Census collection is often viewed with suspicion. Local authorities may suspect that it is to be used for the levying of a national tax and thus there may be an incentive to under report population size. Alternatively it may be suspected that the census is to be used for allocating development funds in which case there may be an incentive to over report.

Conclusions and Summary

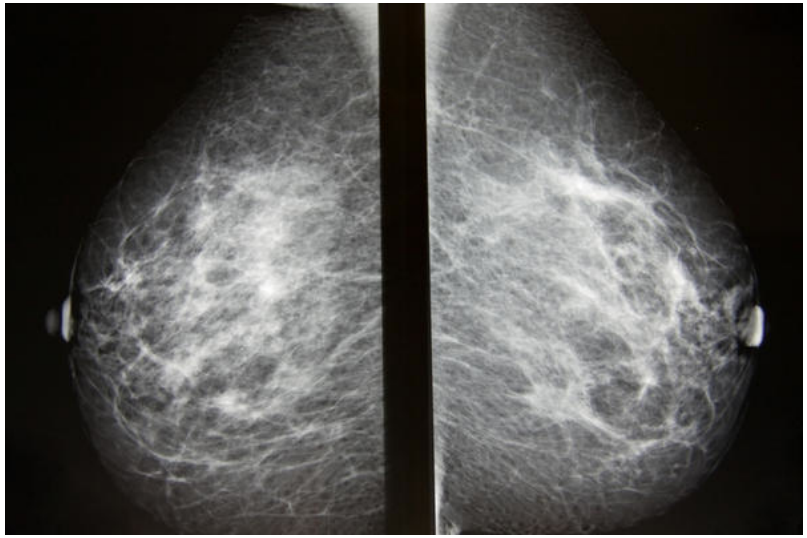
- Whatever the case, although (at face value) the population estimations produced were not as accurate as the "ground truth" census data (this was to be expected), the proposed method offered significant cost and time savings.
- Overall processing times of about 30 minutes was recorded, as opposed to the many days that would be required to conduct the original survey using traditional methods.

Concluding Thoughts on Image Mining

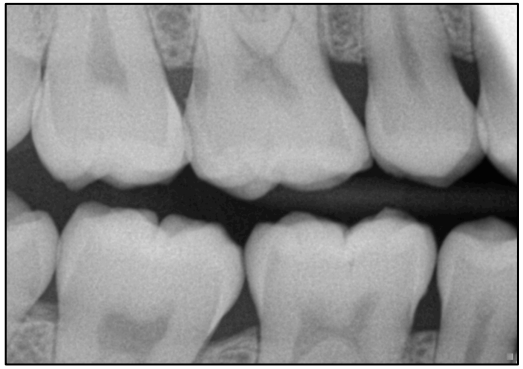
- Decisions are regularly made with the support of imagery of some sort (Satellite Image, MRI, OCT, etc.).
- Our ability to collect imagery of all kinds has enhanced rapidly over the last decade (we can do it cheaper and faster).
- We have seen rapid growth in the global image sensor market.
- Analysis is still often conducted manually, very little software automation (although some support tools do exist, e.g. Brain Voyager for MRI brain scan data).

Further Work

- Still issues with explanation generation.
- Lot of scope for alternative representations, especially fuzzy and deep learning approaches.
- Lots of scope for further application.



(a) Mammogram



(b) Dental X-ray

(c) Hand X-ray