An ESRC Data Investment

ZOOPLA

Delineation of perceptual Neighbourhoods in the London housing market - geo-text location modelling through Natural Language Processing

Consumer Data Research Centre

Author: Vladimir Tesniere | **UCL Geography – MSc Social and Geographic Data Science**

UCL Supervisor: Dr Stephen Law | Partner Institution Supervisors: Dr Chanuki Seresinhe, Dan Mepham-Lagrue (**Zoopla**)

Background and Motivation

Neighbourhoods (NBHD) and their associated areas have meaning to the person on the street, but little information is provided by authoritative data providers in the UK. ***We propose a set of methodologies to construct area delineation****, improving NBHD understanding of spatial and temporal characteristics, whilst also aiming to improve consumer interaction. We use detailed listing text data from Zoopla and Hometrack to run NLP tasks for the creation of these areas. Our most promising proposed methods leverage:*

- **Zero-Shot Encoding** for text classification
- **Topic Modelling** and **NER** for unsupervised clustering

Data and Methods

- Dataset used detailing over ***100k*** unique listings in **H1 2023** from Zoopla/Hometrack database
  - paired with Ordnance Survey (*OS*) UPRN for cross-platform property recognition and comparison
- **Data quality** (high) and **temporal availability** (high)
  - Zoopla historical data available since 1995 – allowing for comparison of outputs over time
- Proof-of-concept in Greater London as highly complex and varied geographical *NBHD* distribution[1]
- Web-scraped dataset (from public sources) containing of all mentioned living areas providing potential *NBHD* locations
- **Pairing with preconstructed geo-demographically precise areas** (LSOA and OA) to finalise boundary appearance

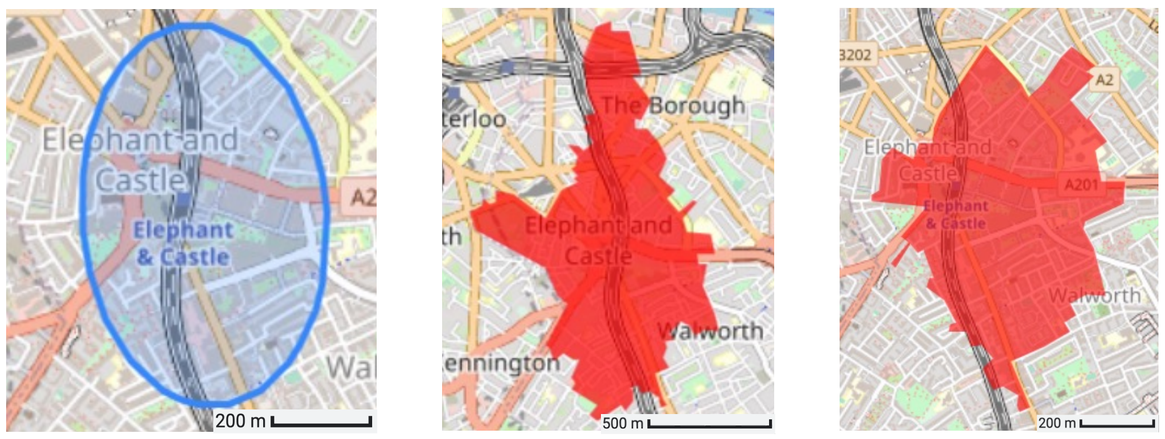


Figure 1: Boundary processing examples for polygon output *(from left to right)* – initial KDE outputs, KDE outputs paired with LSOA, KDE paired with OA

Approaches to difficulties in area delineation

Figure 2: Example of neighbourhood overlap (for 'Kennington', 'Elephant and Castle' and 'Camberwell') and captured through KDE represented boundaries

Figure 1 shows why **defining a *NBHD* can be difficult**, namely considering the following challenges:

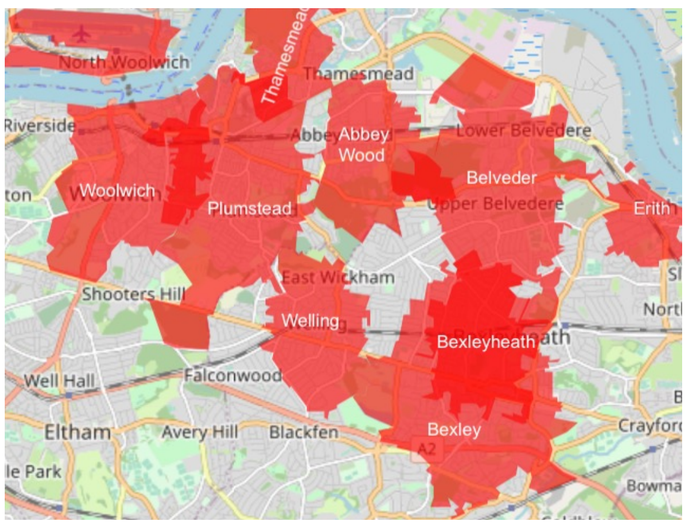
- Boundaries are strongly influenced by the granularity at which we decide to approach the issue

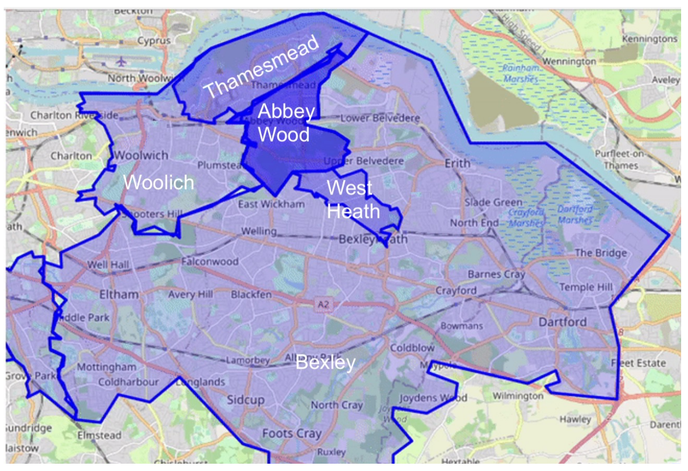
- By nature, these areas **overlap** and are **difficultly discernible**
- **Temporal sensitivity** as these areas evolve depending on population trends and new landmarks
- **No existence of objective definitions** for these areas
- Bias through user preference with areas having 'positive' or 'negative' connotations

Results of Main Model: Zero-Shot Classification

Zero-shot classification aims to associate a text to a label, irrespective of domain

- precise classification of areas through the listing summary description
- Resulting **output displays multiple classes**[2], such that a property can be associated to multiple *NBHDs*, accounting for potential overlap





Considerable improvement in granularity and accuracy on existing sources:

Figure 3 (above): East London Example of our model
Figure 4 (below): current boundaries used by consumers on ZGP LTD

Unsupervised Classification

SBERT pre-trained model for semantic recognition[3] in unsupervised clustering using **HDBSCAN**

- Includes higher potential granularity
  - Over 200 *NBHD* identified (including new area)
- **Less effective** in lower demanded **and more deprived areas** where model has less information

Name Entity Recognition (**NER**) model specific to geo-text

- **Results too superficial**, reflecting that geo-text classifier pre-trained models lack higher granularity area recognition

Figure 5: QR for online interactive map outputs

Validation Approaches

Due to limited ground-truth sources in area delineation, we need forms of validation – **these indicate our boundaries are reliable!**

Table 1: Zero-shot classification Overlap comparison with OS Source (**NB**: **IoU** significant at 0.5<)

Area Name (East London)	Overlap with OS source (%)	Intersection over Union – btw. 0 and 1
Beckton	41.01	**0.57**
Canning town	68.72	**0.50**
Stratford	87.42	0.48
Manor Park	87.96	0.48
East Ham	96.71	**0.54**
Forest Gate	98.14	0.37

For Validation, we propose:

- **Overlap comparison** to manually built boundaries, created from OS data-sources
- Validation through **property characteristic** within the same area (mean m^2 price)
- Comparing to null model, where areas are constructed only on distance weighting

References:

1. Wessendorf, S., 2016. Commonplace diversity and the 'ethos of mixing': perceptions of difference in a London neighbourhood. In *Ethnography, Diversity and Urban Space* (pp. 60-75). Routledge.
2. Yin, W., Hay, J. and Roth, D., 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv preprint arXiv:1909.00161.
3. Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.