Using Visual Analytics to Detect Problems in Datasets Collected From Photo-Sharing Services

Alexander Kachkaev, Jo Wood – giCentre, City University London, UK

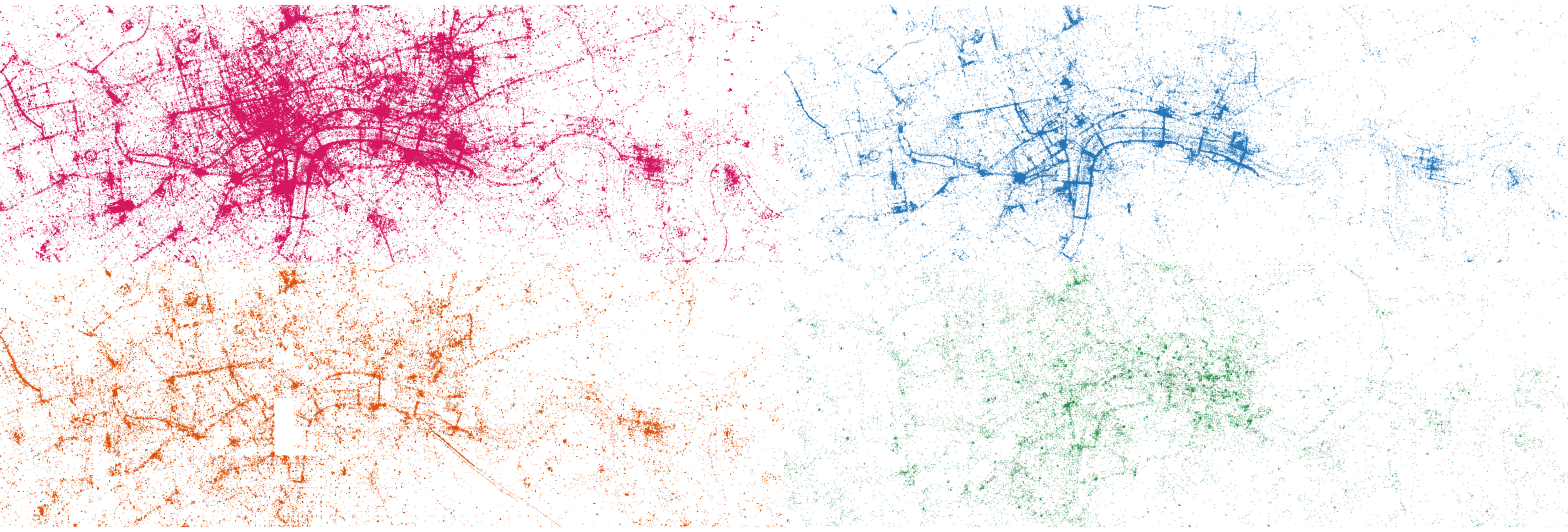
Datasets that are collected for research often contain millions of records and may carry hidden pitfalls that are hard to detect. This work demonstrates how visual analytics can be used for identifying problems in the spatial distribution of crawled photographic data — input for a pedestrian routing system that suggests attractive paths.

Overall Density Evaluation

Showing density with semi-transparent circles was found to be a useful instrument for assessing geographic structure. For instance, it is clear that the street network is well-seen in case of Flickr and Panoramio, while barely apparent in Geograph, making it less suitable for street navigation.

Flickr

Panoramio



Picasa

Geograph

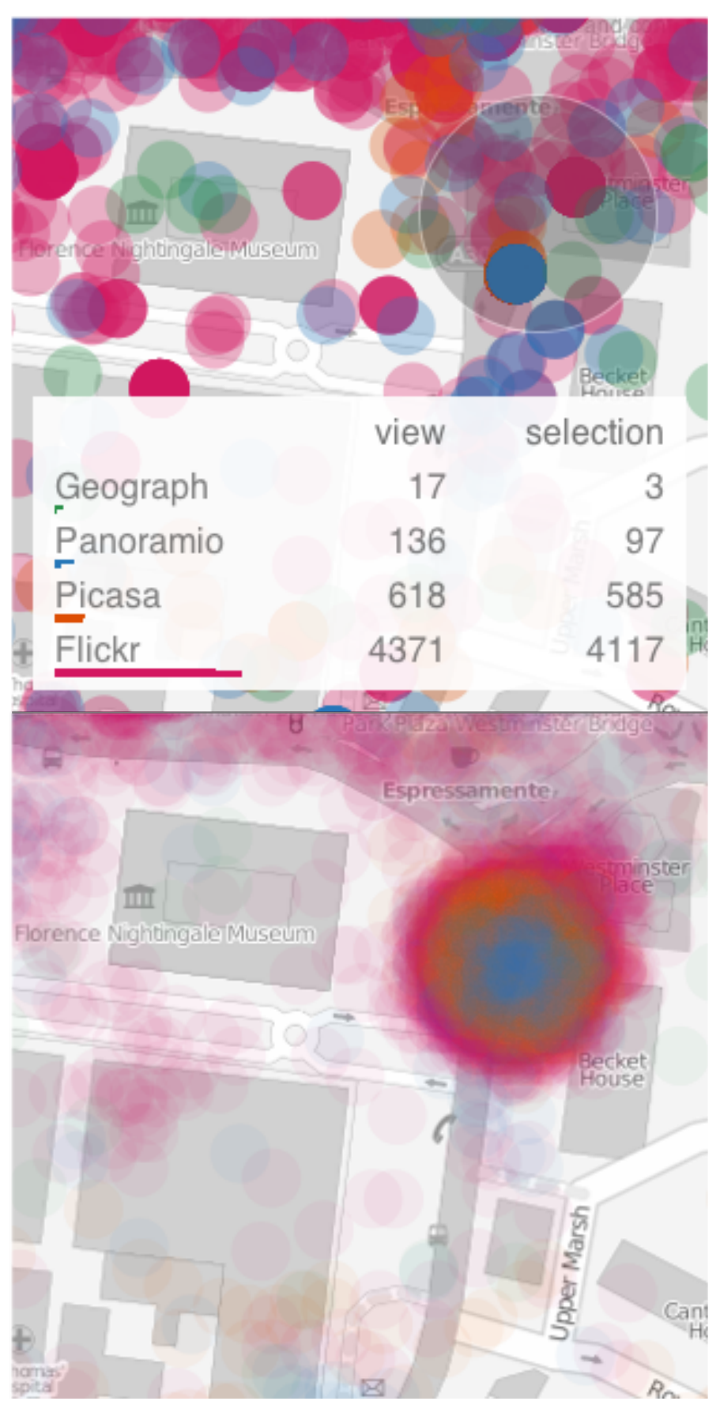
API Failure Detection

Photo service APIs may fail to return all items in a requested region. VA helped to detect and investigate such behaviour in order to improve crawling process or understand why collection of photographic metadata is not possible in a particular case.

Detection of a possible problem with Picasa API and its confirmation by visualising results of a single API request. Photos taken between January 1st , 2008 and December 31st , 2011 are orange, the rest (generally, most recent) are grey. *Bottom:* The server returned photographs outside the requested area.

Misplaced Photograph Detection

Because many photographs are geotagged manually rather than using GPS (in particular by using place search) some spots end up containing hundreds of photographs, not related to the local areas where they are placed. Visualisation helped to find such spots.

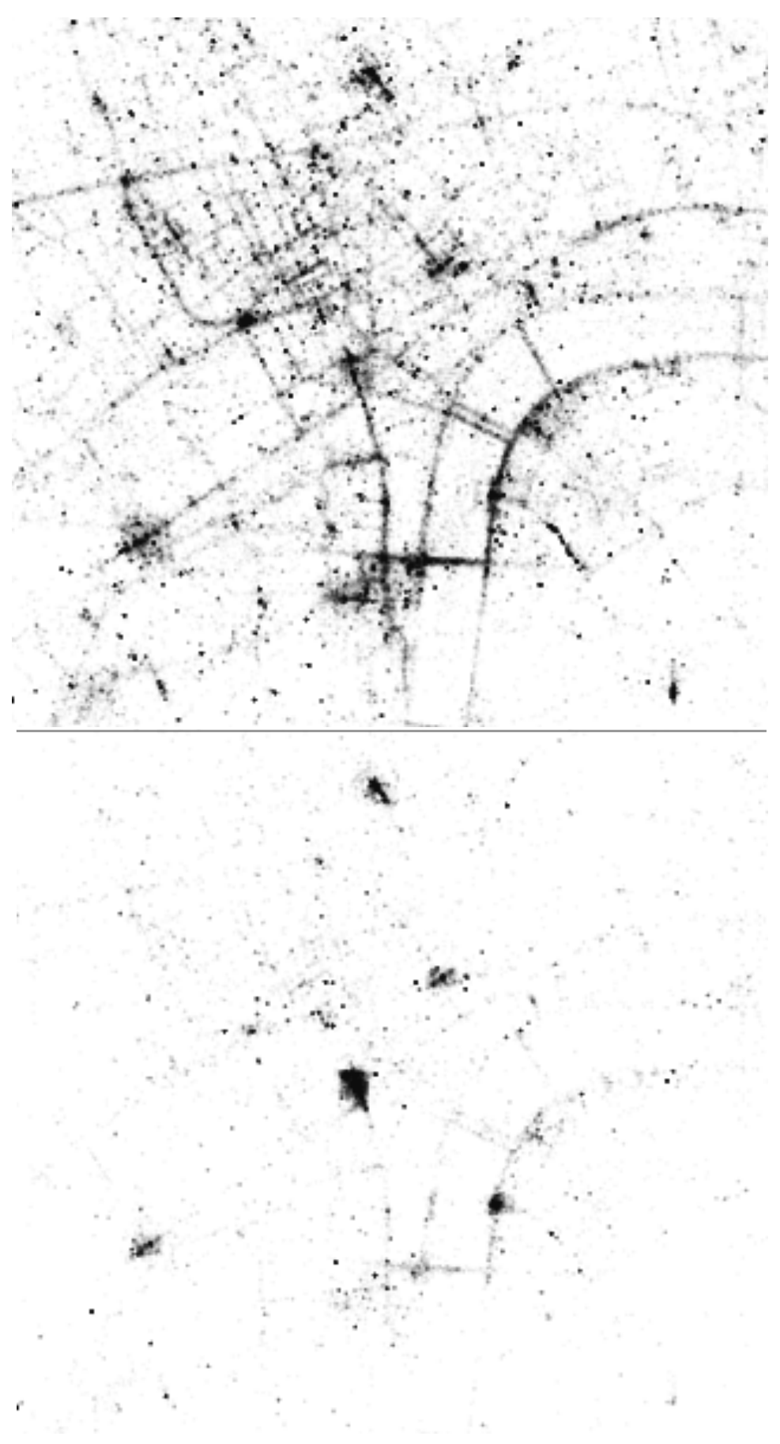


	view	selection
Geograph	17	3
Panoramio	136	97
Picasa	618	585
Flickr	4371	4117

An example of an anomaly containing unexpected numbers of photographs with identical coordinates. Adding a randomly generated offset (standardised to fit a normal distribution) to photo coordinates and reducing alpha helps to find such places. *Background:© CC BY-SA OpenStreetMap and contributors.*

Comparison of Crawling Methods

Visual analytics helped to see that the choice of metadata crawling approach can significantly influence the distribution of the photographs in resulting dataset.



Photographs collected from Flickr using *top:* items only returned when using spatiotemporal requests (with the minimum edge of a bounding box ≈100m) and *bottom:* items only returned when using user-content requests (data kindly provided by Gennady Andrienko, Fraunhofer Institute, Germany).

Other Implications of the VA Approach

- Visualisation of distribution of photo illuminance based on EXIF data to see if a dataset contains significant numbers of photos taken indoors or overnight, which are not wanted for some analysis
- Time filtering in order to observe dynamics in spatial-temporal distribution of the photographs caused by events and seasonal changes
- View of photographs with faces in them, with information obtained by means of service APIs using image processing to see the amount and the distribution of private photographs.

CITY UNIVERSITY
LONDON

http://openaccess.city.ac.uk/1320
alexander.kachkaev.1@city.ac.uk

gicentre.org