# Reconstructing historical rural addresses with VGI and digitized aerial photography

Mads Linnet Perner[1[0000-0003-3890-207X]] and Stig Svenningsen[1[0000-0001-7949-0740]]

[1] Royal Danish Library, Søren Kirkegaards Plads 1, 1221 Copenhagen, Denmark
stsv@kb.dk

This paper describes an attempt to develop a historical GIS of farms with metadata from digitized aerial photography. With the current effort of mass digitization, wide ranges of new data sources are becoming available to historical scholars. However, for these digital sources to be included in new scholarship, a prerequisite is the research infrastructure necessary to process them. Our project has examined how metadata on digitized aerial photography, generated by volunteers in the Royal Danish Library's crowdsourcing effort, might provide a short cut to historical GIS infrastructure, which would otherwise require significant resources to build. As part of the digitization of thousands of aerial photographs of rural properties, the library had the help of volunteers to geo-locate each photograph. As each photograph often represented a single property, the data points and their metadata are representative of a certain address. This paper outlines the steps we took to develop the raw material into a dataset containing locations of historical rural addresses. Based on a pilot study of a limited area, we discuss the quality and accuracy of the data resulting from our approach. We found that the overall quality of data extracted is acceptable compared with the traditional approach of manually plotting in farm-localities by hand.

## 1 Background

Recent years have seen a significant growth in the number of humanities scholars applying GIS methods to their work, a field that has come to be known as the Spatial Humanities [1]. As mapping becomes an increasingly valid tool for historical research, we have seen an increase in the use of conventional GIS data sources such as census returns and tax registers, but also initiatives to apply GIS to new material types, most notably by the geo-parsing of text. This development is promising, but it raises new demands for geospatial research infrastructure for the digital humanities. In Denmark, the national GIS of historical administrative units, DigDag, has been key, but its most detailed geographical unit, the parish, effectively limits the detail of analysis [2]. Urban historians are advantaged by detailed cadastral maps that were often in use very early, in Copenhagen as early as the late 17th century [3]. Rural areas, however, are less accessible in terms of geographical sources to warrant accurate GIS data. When

cadastral maps are available, they often focus on the division of arable land rather than the exact location of farm buildings.

This paper presents an alternative route to a historical spatial research infrastructure for rural settlements. It aims to examine how the metadata produced by aerial photography businesses, operating mostly in the countryside, can help us locate and map rural settlements and their inhabitants. The aerial photography collection held by the Danish Royal Library includes thousands of such oblique photographs depicting single farms, often with information on the farms or owners name, or a cadastral reference. The photographs have been digitized as part of the project *Danmark set fra Luften* and have further been geo-located by volunteers in a crowd-sourcing setup [4]. In this way, our paper taps into the broader wave of Volunteered Geographic Information (VGI)-generated data and its use in the digital humanities scholarship and beyond [5,6].

Potentially, this data offers a chance to create a farmhouse-level GIS. Such a spatial dataset would provide a much-needed spatial reference to the increasing amounts of historical information digitized by archives, libraries and museums, and published in comprehensive digital repositories. Examples in Denmark include census returns digitized by the Danish National Archives [7], in which farm names are the most detailed spatial units for the rural population, and in the newspaper corpus digitized by the Royal Danish Library [8], in which certain farms often feature in job postings and the like. We consider this paper a pilot project, an experiment, to examine to what extent the Royal Danish Library's aerial photograph collection can be used to locate historical rural addresses. This paper focuses on the island of Bornholm, but our approach has the potential to cover rural settlements in all of Denmark. The resulting dataset can serve as infrastructure for traditional spatially minded fields like historical geography and landscape research [9]. Nevertheless, it may also help acknowledge the spatial dimension in conventional historical scholarship.

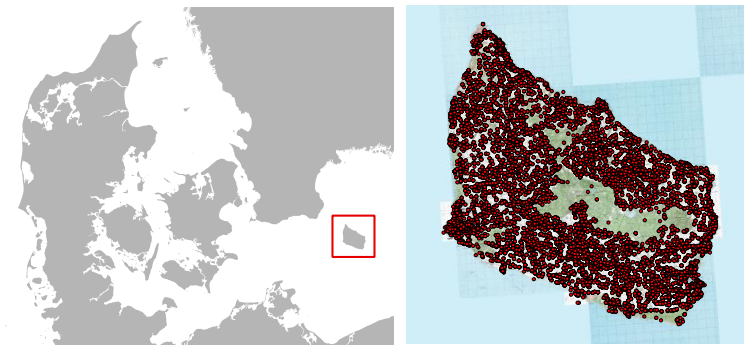

**Fig. 1.** The location of Bornholm in the Baltic Sea (left) and a map showing the distribution of aerial photographs covering Bornholm, which have been assigned coordinates by volunteers, in the library's collection. Sources: Kortforsyningen, Danish Agency for Data Supply and Efficiency and European Environmental Agency (coastline).

## 2 The Aerial Photograph Collection of the Royal Danish Library

The Royal Danish Library holds a collection of more than 5 million aerial photographs covering most of the country from 1890 to 2010. Approximately half of these are oblique images depicting farms and homesteads, that were captured by a handful of aerial photography businesses aiming to sell the pictures as commodities to local land owners [10]. As a result of this venture, in which company salesmen approached customers on their own doorstep, the library's collection contains information on the names of individual farms and their owners, addresses, and other information related to the sales practice in addition to the actual photograph negatives. As such, for the period c. 1930 to 1990 when oblique aerial photographs were common, they can provide us with a means of identifying and locating many rural properties [4].

Starting in 2012, the project *Denmark seen from above* aimed to digitize and publish the Royal Danish Library's collection of aerial photographs to an online portal. In addition to digitization, the photographs are presented in form of a VGI system where visitors are encouraged to contribute by placing photos from their local area on a map. At the time of writing, 1,093,850 of 1,506,723 (72.6%) digitized photographs have been accurately geo-located. For the island on Bornholm, which is the focus of this paper, all of 20,346 photographs covering the island have been geo-located. Thus, if we briefly ignore the actual photographs, what remains is a vector point dataset with, in most cases, either a farm or owner name or an address.

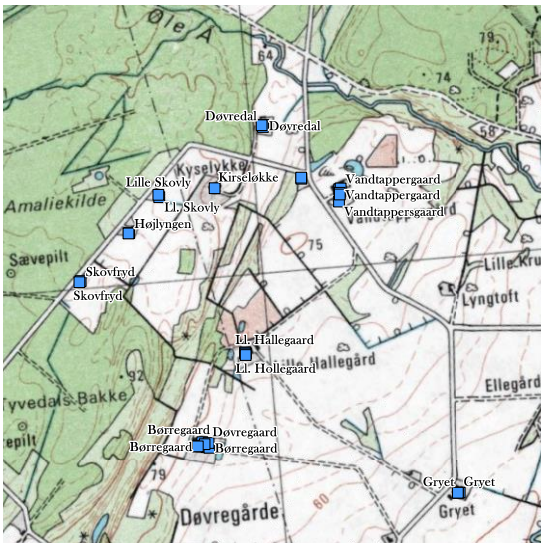


**Fig. 2.** The distribution of aerial photographs in a select area on the south-eastern part of the island. The labels represent each farm name as it is spelled in that particular data entry.

# 3 Extracting rural addresses

### 3.1

For the most part, the metadata of the aerial photographs contains either the name of the property or owner, or an address or cadastral reference. In most cases we only have one piece of information. For example, rarely does the metadata tell us the names of both the property and the owner. As such, for 12,225 photographs (60%) we know the name of the property, commonly a farm name or, in urban areas, shop or factory names. In 3,523 cases (c. 17%) we know the name of the owner. Street addresses are very uncommon in the dataset - they exist for less than 100 photographs in total - so we decided to proceed exclusively with owner and property names.

Figure 2 is indicative of a general pattern, namely that most farms were photographed numerous times over the years. As such, in terms of unique locations, the dataset only numbers around 5,000 places. The clusters of points around each farm are the result of differing practices among the geocoding volunteers; some place the photograph *on* the farm itself, while others have attempted to place the photograph at the point where it was snapped. The figure highlights another central issue with the data: the presence of spelling variations, which makes it difficult to aggregate the points. Abbreviations like 'Ll.', short for 'Lille' (small), are manageable. Other variations are difficult to address in the processing of the data, an example being farm of Lille Hallegård, which, in one case is spelled 'Hollegård'. The farm of Børregård, of which one data point is labeled 'Døvregård', is not a different spelling but rather the name of a neighboring farm not included in the figure. As such, it represents a misunderstanding on the part of the photography business' delegate, which has found its way into the archive, or possibly a misinterpretation made by the metadata typers.

The challenge, then, lay in identifying and linking those groups of data points representing a certain address or farm. In connecting the dots, we can draw on two variables: farm names shared by several data points and geospatial proximity. First, some processing was required to standardize names with slight spelling variations. The data was loaded into OpenRefine, which has clustering function to identify and group similar strings using various text-mining methods. The program suggests groups of similar strings to the user, a feature that allowed us to qualitatively assess whether two strings represented the same name in different spellings or two distinct, but similar names. The clustering method helps to some extent, but there remain several spelling variations which it does not recognize. On the other hand, the sheer size of the data sample rules out manual correction of those variations, and even more so if the project were to be extended beyond the island of Bornholm. Instead, we used short distance between two data points as an indication that those were connected. First, we dissolved data points with the same name into the same object, as such, turning single-part features into multi-part features. Since some farm names are common across the island, the spatial merge was performed using a historical parish delineation dataset. Second, for each set of points with the same name, we calculated the geometric centroid. Figure 3 exemplifies the results of this process. Two of the farms are marked by a single centroid, as all the data points shared the same spelling. The third farm,

'Kattesletsgaard', is marked by two centroids due to an undetected spelling variation. Both points represent the same farm, and since that is the case they are located close to each other. As the distance between farms is usually quite large, around 200 meters on average, we can assume that two points in close proximity, in most cases, represent different ways of spelling the farm name. As such, for data points within 25 meters proximity, we replaced those with their common geometrical centroid.

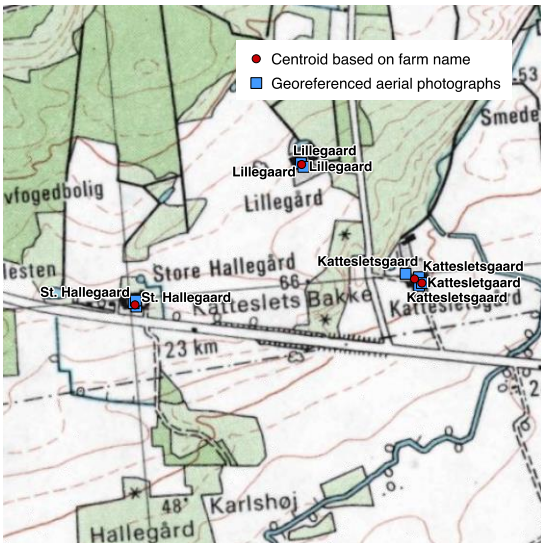


**Fig. 3.** The distribution of aerial photographs of three farms on the southeastern part of the island, along with the centroids generated by our data processing.

### 3.2

The sequence described above produced a dataset with each point representing a farm. The process involves some hand-held work, starting with the raw data extracted from the VGI API, through a number of manipulations in Excel, OpenRefine and ArcGIS. We would argue that this approach is a fast and effective way to create a farm-level database compared to, for instance, hand-held digitization of farm locations from historical map sheets. To evaluate our approach, however, we would need a sample of the latter. We chose a case area within Bornholm: the predominantly rural parish of Østermarie. From the national 1:20.000 topographic map series produced from 1901-1971, we digitized all non-urban buildings with as a point dataset along with the farm name or reference apparent on the map. This data sample allowed us to evaluate the coverage of our dataset built from aerial photographs.

We performed an analysis of the overlap between the two layers: a spatial join linking points from each dataset within 25 meters proximity. Of 247 points in the Østermarie sample, all points were partnered with a point in the air photo dataset. Only

88 points in the sample are accompanied by a name on the map sheets; the remaining farms were too small to be assigned with a farm name on the topographic map. As such, as the analysis displays, we are able to attach names to those farms with the data stemming from aerial photography. In other words, the dataset resulting from our approach has a coverage similar to the hand-held sample but excels in its detail as it contains the names, a key variable for links to census data and other sources, of many of the smallest farms in the area.

## 4 Discussion and conclusion

The results of our pilot study suggest that there is a potential for using the geo-localized metadata produced by volunteers at the crowd-sourcing platform as a data source for establishing a farmhouse-level reference dataset. Such a dataset would be an indispensable resource for realizing the potential of geographical studies of other historical datasets. We can think of several examples where detailed historical addresses can be useful. One is in historical disease modelling, where not only population counts, but also the location of people and data on the geographically unequal distribution of the population is key to studying the spread of communicable diseases, like measles. In landscape studies, our dataset could be linked to census material in order to study how the social status of a farm's inhabitants influenced landscape change in the surrounding area [9]. Finally, in onomastics our farm-level dataset might serve to locate the place names of small settlements that have disappeared and for that reason do not appear on survey maps.

Our success, however, may be conditioned by the high quality of the data from Bornholm. At the time of writing, the digitized aerial photographs have been online in the VGI portal for almost five years. Thus, the volunteers have been working with the data for some time, which is also reflected by the fact that 10% of the photos have been annotated with use comments, ranging from factual information to small anecdotes and reflections, a high share compared to the data for the rest of the country. Further, the amount of metadata typed by library staff has declined during the run of the project. This means that we cannot expect the same quality of metadata going forward when the rest of the country is up for processing. A solution to this could be to mobilize the volunteers in an effort to provide the data needed. This would need to be followed by a communicative effort explaining the need for data. Here, a potential selling point could be the use of such data for genealogical research, which is a strong motivational factor among volunteers participating in crowd-sourcing projects.

To conclude, this pilot study has confirmed that aerial photographs can be an important asset in building a rural historical GIS. However, despite a long-running digitization project and a very effective VGI effort, plenty of work is still require in order to work the aerial photographs into solid geographical reference data, and more time than we have been able to invest in this pilot examination. One aspect of the metadata, which turned out to require much work, was the standardization of names of persons and places. It is worth considering that if our data were to be linked together with, for instance, census records, the names in that dataset would require

standardization as well, if not for some sort of record linkage using text pattern recognition. An encouraging finding of our study, however, is that our aerial photography data contains the place names of many rural settlements that appear on the Danish survey maps, but were deemed too small for their name to be included in print.

## References


1. Gregory, I. N., Geddes, A.: Introduction: From Historical GIS to Spatial Humanities: Deepening Scholarship and Broadening Technology. In: Gregory, I. N., Geddes, A. (eds.) Toward Spatial Humanities. Historical GIS and Spatial History, pp. 35-61. Indiana University Press, Bloomington and Indianapolis (2014).
2. Gammeltoft, P: Historical Geography as Research Infrastructure: A Presentation of DigDag, the Digital Atlas of Denmark's Historical-Administrative Geography. In: 14th International Conference of Historical Geographers, Kyoto 2009, pp. 227-227 (2010).
3. Dam, P.: De digitale matrikelkort over København 1689, 1756, 1806 og 1860. Saxo Institute, Copenhagen (2012).
4. Svenningsen, S. R., Brandt, J., Christensen, A. A., Dahl, M. C., Dupont, H.: Historical oblique aerial photographs as a powerful tool for communicating landscape changes. Land Use Policy 43, 82–95 (2015).
5. Sui, D., Elwood, S., Goodchild, M: Volunteered Geographic Information, the Exaflood, and the Growing Digital Divide. In: Sui, D., Elwood, S., Goodchild, M (eds.) Crowdsourcing Geographic Knowledge. Volunteered Geographic Information (VGI) in Theory and Practice, pp. 1-14. Springer, Dordrecht (2013).
6. Southall, H., Aucott, P., Fleet, C., Pert, T., Stoner, M.: GB1900: Engaging the Public in Very Large Scale Gazetteer Construction from the Ordnance Survey "County Series" 1:10560 Mapping of Great Britain. Journal of Map & Geography Libraries 13:1, 7-28 (2018).
7. The Danish National Archives. Dansk Demografisk Database, http://www.ddd.dda.dk/forskning.htm, last accessed 2018/10/29.
8. The Royal Danish Library. Mediestream - Aviser, http://www2.statsbiblioteket.dk/mediestream/avis, last accessed 2018/10/29.
9. Svenningsen, S., Perner, M. L.: The potential of a digital, transdisciplinary approach to landscape change and urbanization around Copenhagen in the 20th century, Geogr. Tidsskr.-Dan. J. Geogr., pp. 1–8 (2018).
10. Hansen, M. D: Luftfotografiets vej til succes. Magasin fra Det Kongelige Bibliotek 25 (4), 32–38 (2012).