# Developing a Geocomputational Workflow to Check the Consistency of Volunteered Geographic Information

H. Mehdi Poor, R. Zurita-Milla, M.J. Kraak

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente,
PO Box 217, 7500 AE Enschede, The Netherlands
Telephones: +31 534874437
Emails: h.mehdipoor@utwente.nl

## Abstract

Checking the consistency of volunteered geographic information is a challenging task. Current approaches addressing this challenge are costly and time consuming as they heavily rely on human interventions. Here, we propose a geocomputational workflow based on the availability of relevant contextual geoinformation. The workflow consists of three main steps: 1) dimensionality reduction using the t-SNE to facilitate further analysis and interpretation of the results, 2) model-based clustering to group the volunteered observations according to their context conditions, and 3) boxplots to highlight inconsistent volunteered observations for each of the clusters. This workflow was successfully tested using volunteered observations on the timing of the first flower of lilac plants. Results indicate that several observations were inconsistent (i.e. unusually early/late regarding their climatic context). These observations are not necessarily wrong, but they should be used carefully to prevent introducing unduly biases in subsequent scientific studies.

**Keywords:** VGI, consistency, contextual geoinformation, model-based clustering, t-SNE.

## 1. Introduction

Recent improvement in information communication and mobile location-aware technologies has led to the production of huge amounts of geoinformation by volunteers. This phenomenon is called volunteered geographic information or VGI (Goodchild, 2007). In VGI, non-experts collect, distribute and, even, analyse geoinformation. These activities provide scientists with a novel source of data. For instance, VGI is commonly used in phenology, i.e. the study of periodic plant and animal life cycle events and how seasonal and inter-annual variations in climate affect them.

VGI quality is, however, a major concern, especially when applying it for quantitative analyses(Flanagin and Metzger, 2008, Goodchild and Li, 2012). This is because VGI does not often follow scientific principles of sampling design, and levels of expertise vary among volunteers (Kelling et al., 2011, See et al., 2013, Comber et al., 2013). Moreover, unlike traditional geographic information, VGI typically lacks quality checking mechanisms (Elwood et al., 2013, Goodchild and Li, 2012).

Among VGI quality problems, inconsistency is considered an important one since it biases VGI-based analysis and modeling results (Schlieder and Yanenko, 2010, Yanenko and Schlieder, 2012, Ferster and Coops, 2013). Inconsistent VGI are those that are implausible regarding their geographic locations or time. They appear frequently in VGI due to the non-professional and subjective character of VGI.

Current approaches to deal with inconsistency in VGI mostly rely on human interventions. However, such approaches are often costly and time-consuming, and are impracticable in many situations such as monitoring of fast-changing phenomena such as phenological events (e.g. first flowering). Clearly there is a need to identify inconsistent VGI in a robust and automated fashion.

This study proposes an automated workflow for identifying inconsistent VGI that uses contextual geoinformation and computational methods to solve the problem. The rest of the paper describing steps and results of the workflow setup is as follows: materials and methods are briefly reviewed in the data and the workflow steps sections respectively. After that, the results of the workflow test in a real-world phenological case study are presented and discussed in the results and discussion section.

## 2. Data

Volunteered phenological observations were used to test the proposed workflow. We used a dataset with the location, the year and the day of the year of the first flower of cloned lilacs. The geographic extent of this dataset covers the contiguous United States and the observations are available from 1956 to 2013. This dataset was obtained from the USA national phenology network[1] database.

The timing of most phenological events correlates well with climatic parameters. In fact, this is the reason why phenology is becoming one of the most popular methods to evaluate climate change. Therefore here we use the most detailed set of climatic data for the US, namely the DAYMET database[2]. The following cumulative daily climatic parameters were calculated to characterize the environmental context conditions in which the lilac observations took place: daily surface minimum and maximum temperatures, precipitation, humidity, shortwave radiation, snow water equivalent, and daylight. The cumulative period stars the first of January of each year and ends the day of the year in which the flowering was recorded. Lilac observations done before 1980 were not analyzed since this is the first year of the DAYMET database.

## 3. The workflow

The proposed workflow applies a specific sequence of methods and techniques to identify inconsistent observations (Fig. 1). Clustering VGI based on the contextual condition in which they were collected provides considerable information about the variability that one should expect in the VGI dataset. When the contextual information is high-dimensional, mapping it to a low-dimensional space facilitates both the clustering and the subsequent visualization of the results. Once the observations are assigned to clusters, inconsistency is identified by looking at the outliers present in each cluster.

[1] https://www.usanpn.org/results/data

[2] http://daymet.ornl.gov/dataaccess.html

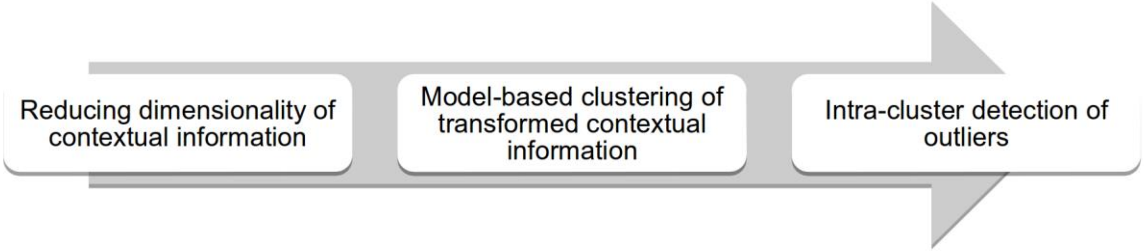
Reducing dimensionality of contextual information
Model-based clustering of transformed contextual information
Intra-cluster detection of outliers

Figure 1. The main steps of the workflow for identifying inconsistent VGI.

First, dimensionality reduction of the contextual geoinformation is needed since we have a large number of contextual variables. We selected the t-distributed stochastic neighbor embedding or t-SNE (Van der Maaten and Hinton, 2008) to reduce the dimensionality of the contextual geoinformation while its local structure is kept. This means that similar objects are mapped to nearby points in the low-dimensional space. The chosen number of final dimensions depends on the complexity of the data. Choosing two dimensions could facilitate the visualization of the clusters in this study. The inclusion of the t-SNE improved the workflow as it allowed the use of a wide range of the contextual geoinformation to create clusters with similar context conditions.

Model-based clustering using normal mixture models (Banfield and Raftery, 1993, Fraley and Raftery, 2002) is used in the proposed workflow as it address following questions in an automated fashion: 1) how many clusters there are 2) what is the shape and size of clusters 3) how should outliers be handled and which distribution model should be used. The automated identification of cluster characteristics is realized by sequentially fitting several mixture models to the transformed contextual geoinformation and selecting the one that maximizes model selection criterion, which here is the Bayesian Information Criterion (Biernacki et al., 2000). However, the efficiency of the selected clustering method might be negatively affected by the dimensionality of the input data, i.e. the number of contextual variables selected to characterize the context condition. This justifies the use of t-SNE method in the preceding step.

The Tukey boxplot (Frigge et al., 1989), a hybrid method that displays variation and outliers in numerical data, is proposed as the final step of the workflow to detect the intra-cluster outliers, i.e. inconsistent observations. Keeping in mind that VGI collected from similar context should follow a normal distribution; boxplots of the VGI belonging to each cluster can quickly detect intra-cluster outliers, an observation with values higher/lower than 1.5 the interquartile range, i.e. distance between third quantile and first quantile of the data in each cluster.

## 4. Results and discussion

The results of model-based clustering and their corresponding uncertainty are shown in Fig. 2. Notice that the clustering was performed on a two dimensional data created by using the t-SNE method. The clustering identified both well-isolated groups and “fuzzy” groups. Clustering uncertainty is high where the clusters are close to each other and mixed. This uncertainty was calculated as the complementary value of the probability of the most likely group for each observation. Observations with an uncertainty higher than 0.5 were ignored from the further step as they could be either an inconsistent or a mis-clustered observation.

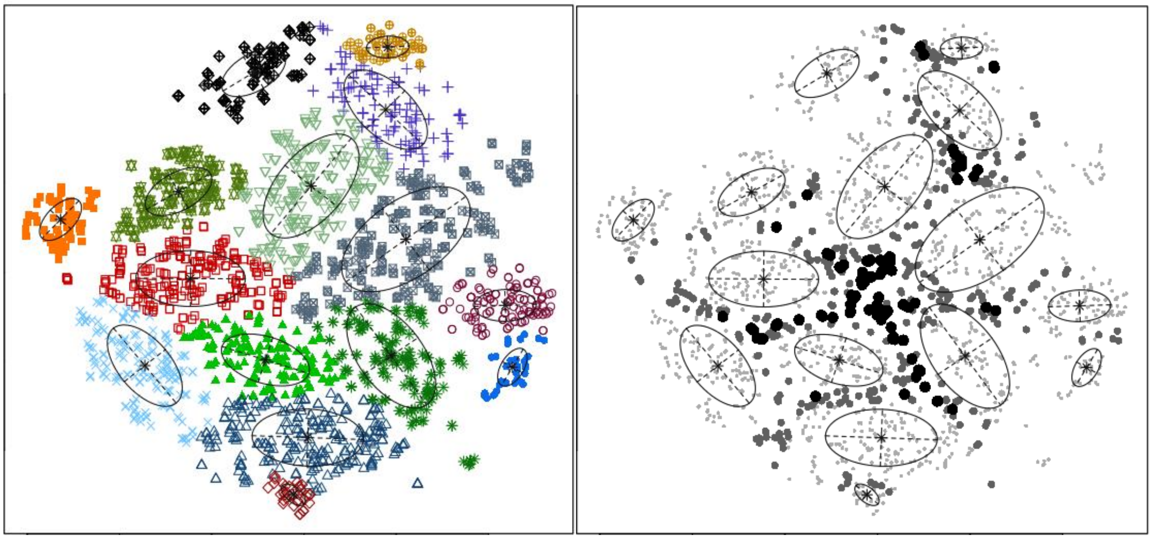

Figure 2. The model-based clustering results in which transformed contextual geoinformation are clustered in different colors (left). The corresponding uncertainty in clustering plot (right).

Fig. 3 shows the geographic distributions of the volunteered observations belonging to each of the 15 clusters identified in Fig. 2. Clusters 4 and 6 contain some spread in the geographic distribution of the observations while the rest of the clusters tend to be compact. Some clusters have a geographical overlap such as clusters 3,11,13 and 14. Such overlaps point out a variety of contextual conditions might exist in northeast of the US.

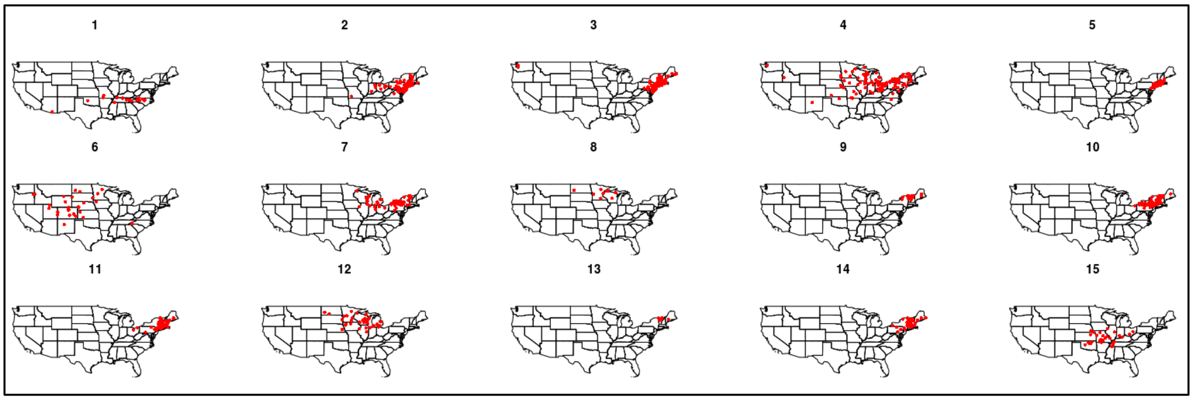

Figure 3. The geographical distribution of clusters through the study area.

The intra-cluster outliers (hollow circles in Fig. 4) founded as inconsistent volunteered observations are presented in Fig. 5. They account for around 3% (69 out of 2296) of the all volunteered observations. They were mostly concentrated in northeast of the US, where three different climatic zones are located close to each other: humid subtropical and humid continental with warm and cool summer.

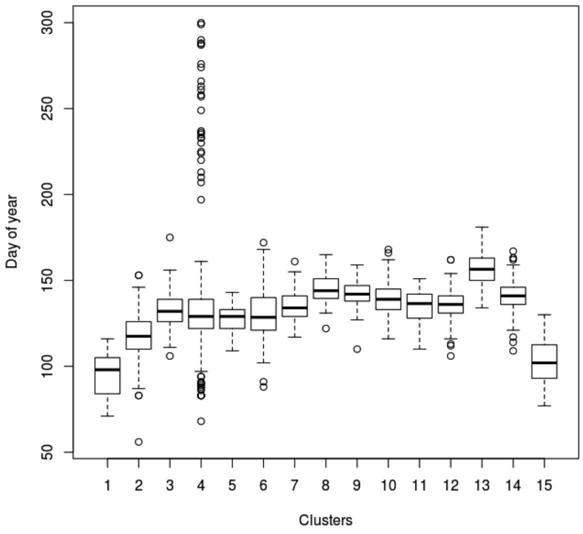

Figure 4. Intra-cluster boxplots of the day of the year that cloned lilac started flowering.

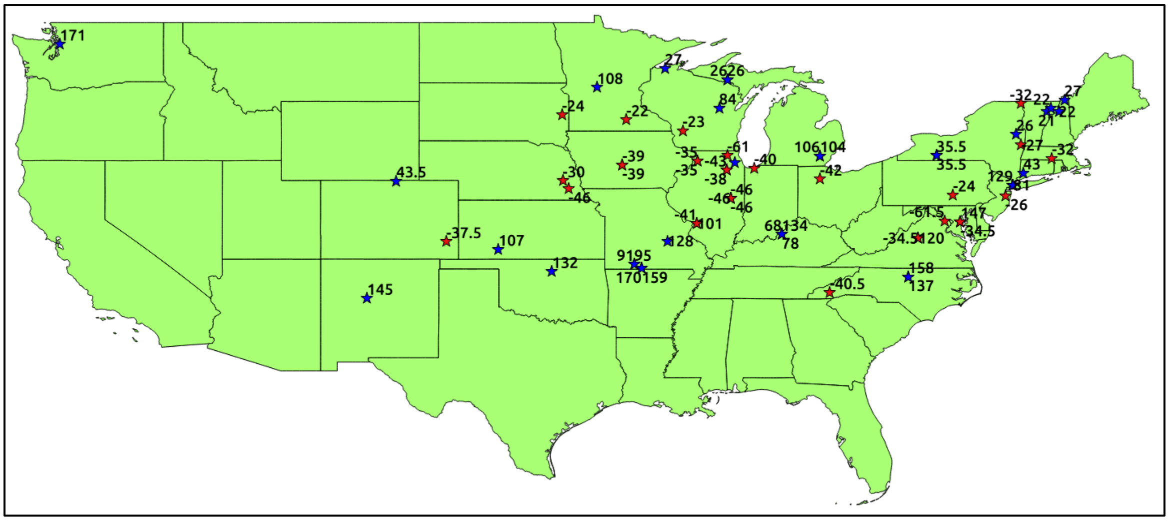

Figure 5. Plot of the inconsistent volunteered observations though study area: The red stars show unusually early while blue ones show unusually late inconsistent observations. The labels show the difference between the day of the year recorded by volunteers and the median of the cluster they belong to.

Double checking with reference geoinformation collected by the US national phenology network, identified inconsistencies could be linked to backyard gardening effects or volunteer records on a weekend or holiday (Fig. 6). This means that inconsistent VGI are not essentially wrong observations. However, they should carefully be used to prevent unduly biases in subsequent scientific studies.

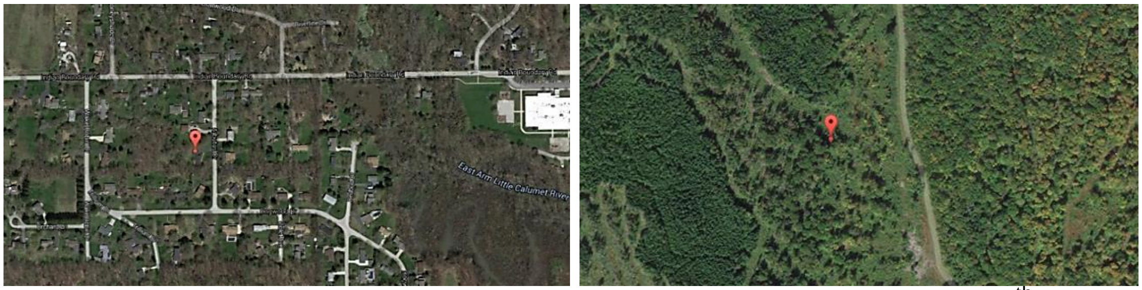

Figure 6. An unusual early first flower at northwest of the Indiana State on 30th March (left). An unusual late first flower at east of the New York State on 8th June (right).

In conclusion, we believe that other VGI initiatives can use our workflow to identify inconsistent observation in phenology but also in other environmental applications. The workflow is based on machine power which clearly makes inconsistency identification less costly and time-consuming. However, the efficiency of the workflow needs to be evaluated in other real-world case studies, which is considered as the perspective of this study.

## 5. Acknowledgements

The authors would like to thank the USA National Phenology Network (https://www.usanpn.org/) for providing the dataset as well as phenological information about lilac flowering.

## 6. References

Banfield, J. D. & Raftery, A. E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803-821.

Biernacki, C., Celeux, G. & Govaert, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22, 719-725.

Comber, A., Brunsdon, C., See, L., Fritz, S. & McCallum, I. 2013. Comparing Expert and Non-expert Conceptualisations of the Land: An Analysis of Crowdsourced Land Cover Data. *In:* Tenbrink, T., Stell, J., Galton, A. & Wood, Z. (eds.) *Spatial Information Theory*. Springer International Publishing.

Elwood, S., Goodchild, M. & Sui, D. 2013. Prospects for VGI Research and the Emerging Fourth Paradigm. *In:* Sui, D., Elwood, S. & Goodchild, M. (eds.) *Crowdsourcing Geographic Knowledge*. Springer Netherlands.

Ferster, C. J. & Coops, N. C. 2013. A review of earth observation using mobile personal communication devices. *Computers & Geosciences*, 51, 339-349.

Flanagin, A. J. & Metzger, M. J. 2008. The credibility of volunteered geographic information. *GeoJournal*, 72, 137-148.

Fraley, C. & Raftery, A. E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Frigge, M., Hoaglin, D. C. & Iglewicz, B. 1989. Some Implementations of the Boxplot. *The American Statistician*, 43, 50-54.

Goodchild, M. F. 2007. Citizens as sensors: web 2.0 and the volunteering of geographic information. *GeoFocus*, 7, 8-10.

Goodchild, M. F. & Li, L. 2012. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1, 110-120.

Kelling, S., Yu, J., Gerbracht, J. & Wong, W. K. Emergent Filters: Automated Data Verification in a Large-scale Citizen Science Project. e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on, 2011. IEEE, 20-27.

Schlieder, C. & Yanenko, O. Spatio-temporal proximity and social distance: a confirmation framework for social reporting. 2010. ACM, 60-67.

See, L., Comber, A., Salk, C., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F. & Obersteiner, M. 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS one,* 8**,** e69958.

Van der Maaten, L. & Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research,* 9**,** 2580-2605.

Yanenko, O. & Schlieder, C. 2012. Enhancing the Quality of Volunteered Geographic Information: A Constraint-Based Approach. *Bridging the Geographic Information Sciences.* Springer.