International Journal of Biometeorology (2020) 64:1825–1833
https://doi.org/10.1007/s00484-020-01968-z

ORIGINAL PAPER

Outlier detection methods to improve the quality of citizen science data

Jennifer S. Li[1] · Andreas Hamann[1] · Elisabeth Beaubien[1]

Received: 11 November 2019 / Revised: 3 July 2020 / Accepted: 6 July 2020 / Published online: 15 July 2020

Abstract
Citizen science involves public participation in research, usually through volunteer observation and reporting. Data collected by citizen scientists are a valuable resource in many fields of research that require long-term observations at large geographic scales. However, such data may be perceived as less accurate than those collected by trained professionals. Here, we analyze the quality of data from a plant phenology network, which tracks biological response to climate change. We apply five algorithms designed to detect outlier observations or inconsistent observers. These methods rely on different quantitative approaches, including residuals of linear models, correlations among observers, deviations from multivariate clusters, and percentile-based outlier removal. We evaluated these methods by comparing the resulting cleaned datasets in terms of time series means, spatial data coverage, and spatial autocorrelations after outlier removal. Spatial autocorrelations were used to determine the efficacy of outlier removal, as they are expected to increase if outliers and inconsistent observations are successfully removed. All data cleaning methods resulted in better Moran's *I* autocorrelation statistics, with percentile-based outlier removal and the clustering method showing the greatest improvement. Methods based on residual analysis of linear models had the strongest impact on the final bloom time mean estimates, but were among the weakest based on autocorrelation analysis. Removing entire sets of observations from potentially unreliable observers proved least effective. In conclusion, percentile-based outlier removal emerges as a simple and effective method to improve reliability of citizen science phenology observations.

Keywords Citizen science · Data cleaning · Outlier detection · Data management · Plant phenology · Climate change

Introduction

Citizen science is broadly defined as scientific inquiry that includes volunteers for data collection and/or processing (Silvertown 2009). Citizen science has been documented as early as 3500 years ago with citizens and officials recording locust outbreaks in China (Miller-Rushing et al. 2012). Today, volunteer observers contribute to various research fields, including conservation science, population ecology, environmental risk assessments, pollution detection, and monitoring of the environment to detect change (e.g., Bonney et al. 2009; Silvertown 2009; Dickinson et al. 2012). Citizen scientists enable large-scale scientific data collection that would otherwise not be possible. In general, any type of biological or environmental monitoring over large geographic areas or long time periods tends to benefit from citizen science networks. For example, citizen science-driven projects have been used to identify pollution sources (McKinley et al. 2017). The establishment, spread, and control of invasive species are regularly supported by volunteer observation networks (Crall et al. 2015). In conservation biology, rare plant populations are monitored, and potential threats to populations have been identified through data collected by volunteers (Havens et al. 2012; Vander Stelt et al. 2017).

In the context of environmental monitoring, an important citizen science contribution is the collection of compelling evidence for biological response to global climate change, for example through observing plant phenology, i.e., the seasonal timing of life cycle events (Rathcke and Lacey 1985). Plant phenology programs supported by citizen scientists include the USA National Phenology Network, which monitors the timing of flowering and leafing of approximately 878 plant

✉ Jennifer S. Li
jsli@ualberta.ca

[1] Department of Renewable Resources, Faculty of Agricultural, Life, and Environmental Sciences, University of Alberta, 751 General Services Building, Edmonton, AB T6G 2H1, Canada

species (USANPN 2017), and coordinates broader cooperation of phenology networks throughout the world (Global Alliance of Phenological Observation Networks, https://www.usanpn.org/partner/gapon). In recent decades, data from such phenology monitoring networks have emerged from relative obscurity to the forefront of environmental monitoring. For example, an analysis of published studies in plant phenology within the *International Journal of Biometeorology* alone has increased from approximately 350 papers per decade between 1957 and 2007 to over 1000 contributions between 2007 and 2016 (Donnelly and Yu 2017).

While the use of citizen science data in environmental monitoring and other applications is well established and widespread, questions have been raised regarding the reliability and objectivity of citizen scientists' data, which has led to some programs reverting to the use of professional scientists, or to limiting volunteer involvement (Silvertown et al. 2013; MacKenzie et al. 2017). Lack of expertise, limited training, and lack of commitment by volunteer observers have been cited as potential issues that may compromise the quality and completeness of data records generated by citizen science networks (e.g., Foster-Smith and Evans 2003; Hunter et al. 2013). However, reviewing the literature on validation and quality assessments of citizen science data, Danielsen et al. (2014) and Kosmala et al. (2016) conclude that citizen science data compare favorably with data collected by professionals, given appropriate training, spot checks to validate observations, and statistical analysis to detect anomalies and outliers.

Statistical outlier detection or other data pre-processing are, in fact, common scientific approaches that do not only benefit volunteer data, but are routinely applied to any professionally collected data or instrument measurements that are prone to inaccuracies (DataONE 2017). Here, we evaluate different algorithms designed to detect outlier observations and inconsistent observers from a citizen science network that monitors the timing of bloom for 30 plant species in Alberta, Canada. The Alberta PlantWatch program is one of the longest currently running citizen-science observation networks in North America (Schwartz et al. 2013). The data has been used to document the impact of climate change at northern latitudes (Beaubien and Freeland 2000; Beaubien and Hamann 2011a). Data from Beaubien and Freeland 2000 has been featured in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change as evidence for the impacts of climate change (IPCC 2007).

The network consists of about 700 observers that have reported more than 57,000 bloom dates from 1987 to 2016. Observers are trained through organized nature walks, booklets, and online resources that provide illustrated guides for identification of species and bloom phases (http://plantwatch.naturealberta.ca). Data are recorded and submitted by observers on data entry sheets, which are then manually transcribed to an electronic database, with both steps potentially leading to accidental errors. Additional errors may result from incorrect species identification or misclassification of bloom phases and leaf-out phases (Beaubien and Hamann 2011b; Crall et al. 2011; Fuccillo et al. 2014). As typical for phenology networks, the data covers a large geographic extent and a variety of climates and ecosystems, which makes outlier detection challenging.

One approach to assess data accuracy is to evaluate the internal consistency of observations, based on the general principle expressed by Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Plant blooming and budburst during the spring in temperate regions are predominantly driven by spring temperatures (e.g., reviewed by Rathcke and Lacey 1985), which is inherently related to location and elevation and highly spatially autocorrelated (Schwartz et al. 2014). This study contributes a comparison of five data cleaning methods that rely on different quantitative approaches, including residuals of linear models, correlations among observers, and deviations from multivariate clusters, and percentile-based outlier removal. As a metric for determining the best method for detecting unreliable observations, we evaluate improvements to spatial autocorrelation in accordance with Tobler's first law of geography.

Materials and methods

Study area and data

The study area encompasses the province of Alberta, Canada, and is bound by the 49th and 60th parallel, the 110th meridian to the east, and the 120th meridian and the Rocky Mountains to the west. Alberta is divided into six natural regions and 21 natural subregions based on topography, climate, vegetation, and soils (Fig. 1), which includes grasslands, montane forests, and boreal and sub-boreal forests, as well as alpine and arctic tundra (Natural Regions Committee 2006). We use this landscape classification system to delineate areas that are expected to have similar phenological responses.

Phenological data from the Alberta PlantWatch database covers observations from 1987 to 2016 with a total of 57,745 observations for 30 species. Observation locations are primarily where human population is greater i.e., within areas around large cities and extensive road networks (Fig. 1). We selected four species that are wide-ranging across Alberta, and phenological phases that had long time series of data available for observation locations. The species used in this study were a common deciduous tree, aspen (*Populus tremuloides* Michx.), the herbaceous species: early blue violet

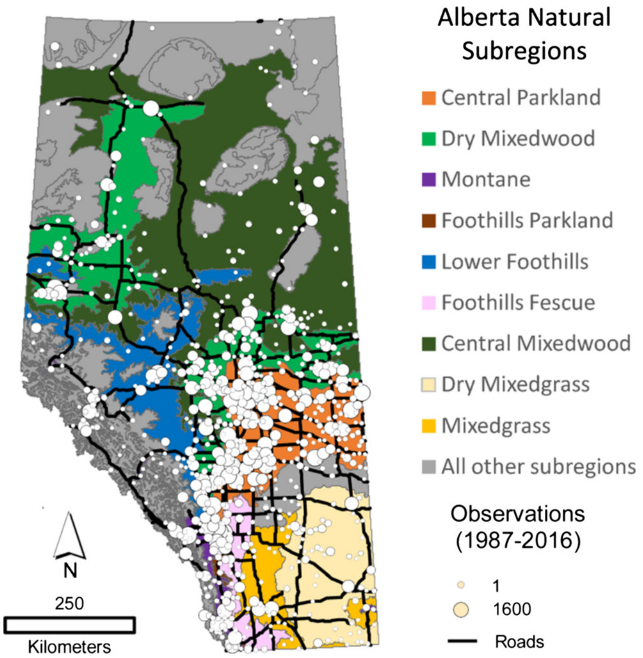

Fig. 1 Observations made by citizen scientists for the Alberta PlantWatch program from 1987 to 2016. The size of the circles represents the number of observations from a single location over this time period

(*Viola adunca* J.E. Smith), and two shrub species: chokecherry (*Prunus virginiana* L.) and saskatoon (*Amelanchier alnifolia* Nutt.). For these species, "first bloom" is reported when the first flowers are open, or when male catkins or cones first start shedding pollen. In addition, "mid bloom" was reported either when 50% of flowers are open or when 50% of male catkins or cones are shedding pollen (Beaubien and Hamann 2011b).

For one of the outlier detection techniques (Mehdipoor et al. 2015), we make use of daily climate data. For this purpose, 1-km resolution gridded climatic data was obtained from the National Aeronautics and Space Administration, DAYMET project for each phenology observation point (Thornton et al. 2016; Hufkens 2017). Following Mehdipoor et al.'s (2015) methodology, climatic variables were summed from January 1 up to the day of the year that each phenological observation was made, and included the cumulative maximum daily temperature, cumulative minimum daily temperature, cumulative average daily temperature (calculated by the average of the maximum and minimum temperatures), cumulative daily day length, cumulative daily precipitation, cumulative daily solar radiation, cumulative daily snow water equivalent, and cumulative daily water vapor pressure.

Data cleaning methods

Five data cleaning methods for improvement of quality for phenological data were evaluated, where each of the methods was allowed to remove 5% of data points:

Method 1 – Standardized difference: Natural subregions generally have similar environmental conditions, so that the expectation is that observations of the same species flowering stage are expected to occur at approximately the same time within nearby areas. A good indication for a potential error would therefore be the deviation of an observation from the mean value of flowering dates for natural subregion, years, and flowering phases. In order to compare outliers across different natural subregions, the scale of observations was standardized to express each observation in units of standard deviations from the natural subregion mean (Fig. 2a).

Method 2 – Linear model 1 with geographic coordinates: This method of data cleaning is based on developing a linear model for the purposes of data cleaning. We follow the methodology of Ranjitkar (2013), where flowering dates were predicted with a linear model based on the continuous predictor variable's latitude, longitude, and elevation. Here, a categorical variable "year" was included as a predictor class variable to account for interannual climatic variability. The records for removal were assessed by calculating the residual difference between the predicted and observed day of the year of the phenological event (Fig. 2b).

Method 3 – Linear model 2 with natural subregion: This is a combination of methods 1 and 2, where natural subregion and year are used as predictor class variable in a linear model for the removal of outliers based on residual error.

Method 4 – Observer correlation: This method targets inconsistent or unreliable observers by identifying low correlations of data reported by individual observers versus the mean of all other observers in the same region (Fig. 2c). Correlations were calculated within subsets for different species, bloom phases, and natural subregions.

Method 5 – Dimensionality reduction and clustering: This method includes contextual environmental information to identify inconsistencies in phenology data. We follow Mehdipoor et al.'s (2015) method, where phenology data, latitude, longitude, elevation, and climatic variables and a number of cumulative daily climate variables were ordinated in fewer dimensions. Briefly, dimensionality reduction was implemented with *tsne* package (van der Maaten and Hinton 2008; Donaldson 2012) for the R programming environment (R Development Core Team 2018). The ordination was conducted by optimizing the Bayesian information criterion (BIC) within the *mclust* package in R (Fraley et al. 2012). Outliers were removed based on the Euclidean distance in reduced dimensions of each observation from the center of its respective cluster (Fig. 2d).

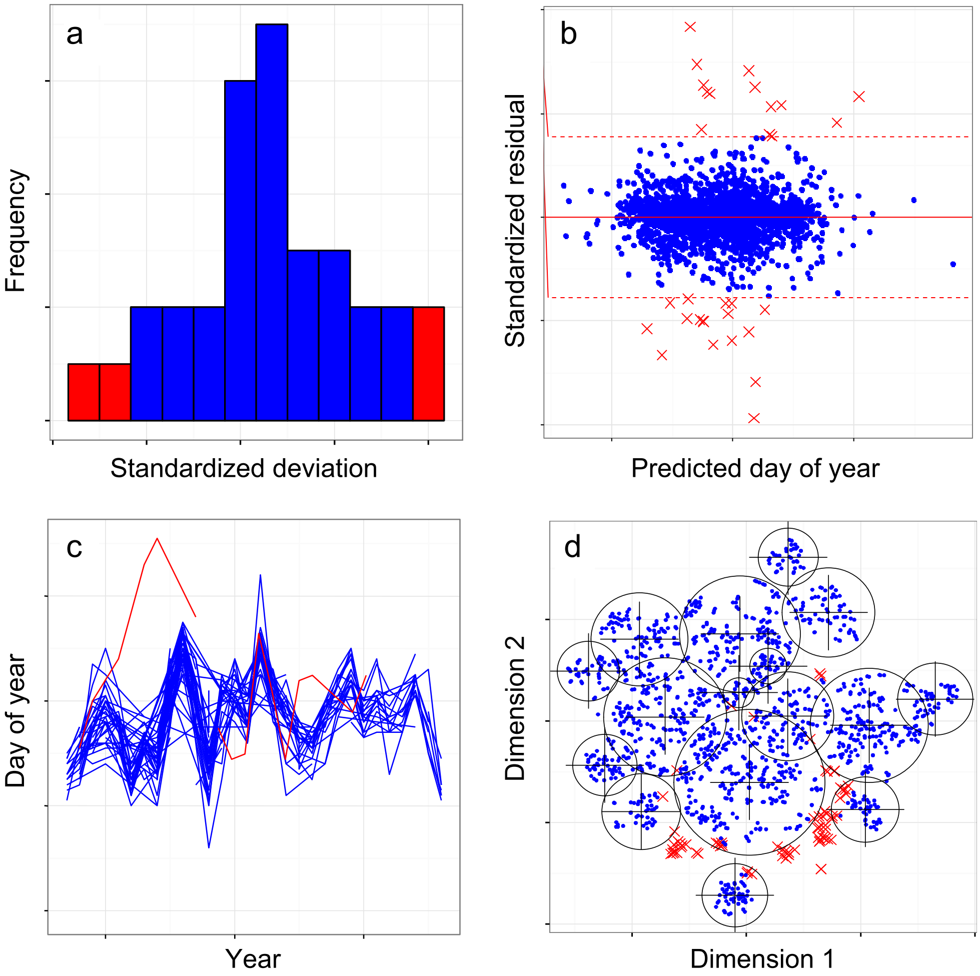

Fig. 2 Visualization of how outlier detection methods work in principle, with the 5% of data that received the highest scores as potentially being an outlier highlighted in red, and the 95% of data remaining indicated in blue. **a** Method 1 – standardized difference. **b** Methods 2 and 3 – linear model residuals. **c** Method 4 – observer correlations. **d** Method 5 – dimensionality reduction and clustering

Evaluation of data cleaning methods

The degree of improvement in data cleaning provided by the five cleaning methods was quantified by Moran's I (Moran 1950) implemented using the *ape* package (Paradis et al. 2004) for the R programming environment. Moran's I ranges from − 1 to 1, where 0 indicates no spatial autocorrelations, positive values indicate positive spatial autocorrelation (i.e., nearby observations are similar), and negative values indicate negative spatial autocorrelation (nearby observations show stronger differences than expected by random chance). Moran's I statistics were calculated for data subsets by year, species, and phase and then averaged. In order to test if improvements in Moran's I values among the data cleaning methods differed among species or bloom phases, we tested for interaction effects among the model effects "method and species" and "method and bloom phase" with a linear mixed model. The effects year, species, and phase were specified as fixed effects, and year was specified as a random effect. The model was implemented by the *asreml* package (Butler et al. 2009) for the R programming environment (R Development Core Team 2018).

In addition to the effect of different cleaning methods on spatial autocorrelations, the effects of the data cleaning approaches were assessed by comparing regional means of the resulting time series, and through comparing maps of observations that were removed by different methods. Regional means by year were estimated with a linear mixed model using the best linear unbiased prediction function of the *asreml* package (Butler et al. 2009) for the R programming environment (R Development Core Team 2018). In this model, the bloom phase was set as the predictor, and year, species, and natural subregion were random effects. As a final visual evaluation of the five data cleaning methods, detected outliers were mapped and examined for similarities and differences across outlier detection methods.

Results

All data cleaning methods resulted an increase in Moran's I when compared with the full datasets prior to data cleaning (Fig. 3). The two methods with the greatest increase in Moran's I value were method 1 (standardized difference) and method 5 (dimensionality reduction and clustering). They were followed by method 2 (linear model with geographic predictor variables), method 3 (linear model with natural regions), and method 4 (observer correlations). The degree of improvement in Moran's *I* varied among species and bloom phases, as indicated by error bars in Fig. 3.

A statistical analysis did not yield significant main effects or significant interactions for Moran's *I* values between the five data cleaning methods and species and bloom phases at an α-level of 0.05, indicating that there is no strong evidence that a particular method is working better overall, or that any method is preferable for a specific bloom phase or species subset. Our sample size is small (four species and three bloom phases) relative to the variability in how the methods affect different species and their bloom phases. Nevertheless, we can state that a difference as large or larger than the observed differences between the control and the two best methods (Fig. 3) would only be expected due to random chance with a probability of 0.11 (method 1, standardized difference) and 0.12 (dimensionality reduction) if the statistical null hypothesis was true.

Regional mean dates showed relatively small shifts in mean dates after the five data cleaning methods have been applied. Ninety-five percent of all mean date estimates for species-region-year combinations shift by 1.1–4.4 days, depending on the cleaning method (Table 1). The largest changes in mean phenology dates (both the 95th percentiles as well as maximum and minimum changes) occurred after two linear models method 2 (linear model with geographic predictor variables) and method 3 (linear model with natural regions) were applied (Table 1). However, with the linear models, the standard error of the phenological occurrences within natural subregions was most effectively reduced, while the other methods resulted in increases of standard errors of the estimates within natural subregions.

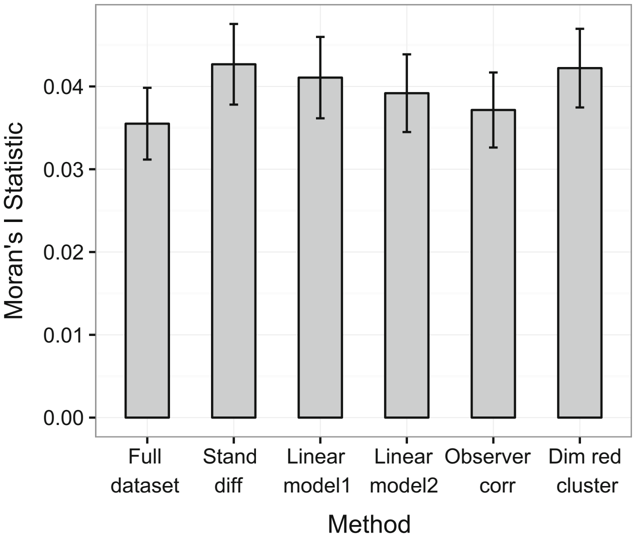

Fig. 3 Effectiveness of data cleaning methods as measured by Moran's *I* statistic (± SE) before and after data cleaning. $N = 240$ groups of datasets per cleaning method (4 species × 2 phases × 30 years). A total of 5% of the records were removed during data cleaning for each data cleaning method. The "full dataset" represents the Moran's *I* statistic for the original records prior to data cleaning

Records that were removed with each data cleaning method were mapped for evaluating differences in spatial patterns of observations that were removed (Fig. 4). The degree of certainty of outlier detection is represented by the size of the circles in Fig. 4. According to the data cleaning method, the metric to scale the circles was the standardized difference, the magnitude of the residuals, the inverse of the correlation to the time series mean, or the linear distance to a cluster mean. While all data cleaning methods removed records in high-density areas (in the south-central region of Alberta), the influence of data cleaning methods varied in low-density areas, in particular the northeast and southeast portions of the province. Method 1 (standardized difference) and method 4 (observer correlation) appear to remove more records in high-density observation areas in south-central Alberta, and retain records in low sample density areas in both the north and the western mountains when compared with any of the other cleaning methods. Method 5 (dimensionality reduction and clustering) appeared to remove more records in the northeast portion of the province. Methods 2 and 3 (linear models) and 4 (observer correlation) did not appear to preferentially remove records in high or low-density sample areas.

Discussion

True validation of citizen science data is usually not possible, unless experts carry out spot checks at the time when data is collected and recorded by volunteers (e.g., Foster-Smith and Evans 2003; Feldman et al. 2018). Generally, citizen science data compares well with professional data given appropriate training, spot checks, and statistical analysis to detect anomalies and outliers (Danielsen et al. 2014; Kosmala et al. 2016). Our study conforms to these assessments, given the relatively small changes to mean observed versus predicted phenological occurrence dates.

While the removal of an individual outlying observation can yield a large change to a particular mean for a species-region-year combination (Table 1), large corrections are quite rare, and 95% of corrections due to data cleaning are small in magnitude (Table 1). This result is similar to observations by Mehdipoor et al. (2015), who worked with a comparable phenology dataset. In their analysis, 97% of the observations were flagged as consistent, indicating that volunteers generally provided reliable information. Yet, if sample sizes for particular species-region-year combination are small, data cleaning can yield substantial corrections, as observed here and also by

Table 1 Change in predicted date for phenological occurrence date for natural subregions after data cleaning compared with the original predicted date for phenological occurrence

Method	Maximum change (days)[1]		95th percentile of magnitude of change (days)[2]	Change in standard error of the estimate compared with control (days)
	Earliest	Latest		
Method 1 (standardized difference)	− 6.6	4.7	± 1.5	0.29
Method 2 (linear model)	− 13.1	15.2	± 4.4	− 0.92
Method 3 (linear model by natural subregion)	− 13.0	14.7	± 4.0	− 0.84
Method 4 (observer correlation)	− 4.3	5.9	± 1.1	0.16
Method 5 (dimensionality reduction and clustering)	− 6.2	6.7	± 2.0	0.16

[1] Maximum change represents the largest changes observed in predicted phenological occurrence date. Best linear unbiased prediction was used to predict the phenological occurrence date in each natural subregion using the *asreml* package in R (Butler et al. 2009). Phase was the predictor, and year, species, and natural subregion were random effects

[2] Difference in days to the mean, estimated using two standard deviations of the differences between the original and post-cleaning occurrence date

Mehdipoor et al. (2015), where the apparent rate of change of bloom dates was corrected by up to 2 days per decade.

After the citizen science observations have been recorded and the data has been digitized, the only indications of problematic data entries are inconsistencies with nearby observations. Some unusual records may actually be accurate, resulting from local climate anomalies that occur in some years, variation in microclimate, or natural genetic variation in the date of flowering among plants. The goal of data cleaning is to remove observations that would exceed this natural statistical variability. Potential sources of inconsistencies in the Alberta PlantWatch dataset may be incorrect identification of plants, incorrect assessment of bloom phases, and data entry and transcription errors, either at the time of observation or at the time of database entry, and other factors (Beaubien and Johnson 1994). The frequency of such errors may vary with volunteer training and experience of the observers. The dataset has also strong spatial biases with many more observations near denser human habitation in the south-central part of the province. This is not an error per se, but

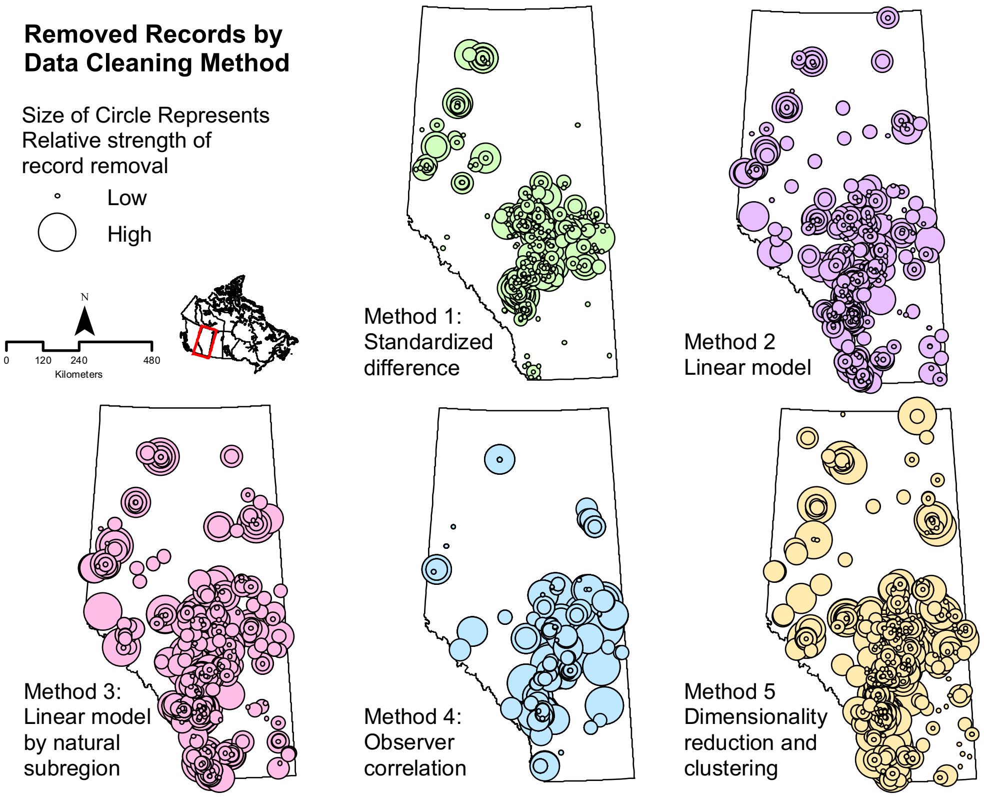
Removed Records by Data Cleaning Method
Size of Circle Represents Relative strength of record removal
Low
High
N
0
120
240
480
Kilometers
Method 1: Standardized difference
Method 2 Linear model
Method 3: Linear model by natural subregion
Method 4: Observer correlation
Method 5 Dimensionality reduction and clustering

Fig. 4 Spatial locations of removed records for assessed data cleaning methods. The size of the circle represents the confidence that the removed observation is an outlier. For method 1, circle size is the standardized difference. For methods 2 and 3, the size of the circle represents the magnitude of the residuals. For method 4, the size of the circle represents the inverse of the correlation to the time series mean, and for method 5, the size indicates the distance to the nearest cluster mean

uneven sampling patterns may still have an effect on how data cleaning methods perform.

Three of the five methods are expected to be unaffected by spatial sampling bias, because they make use of the natural subregion classification system, which captures sampling density well. For example, the Central Parkland natural subregion is a prime agricultural area, where many towns and consequently a high observer density can be found (Fig. 1). Method 1 (standardized difference) and method 4 (observer correlation) made the most direct use of this spatial variable through calculating outlier statistics by natural subregions, before removal of the 5% with the highest outlier scores from the total dataset. In method 3 (linear model 2), one of the model terms was natural subregions.

As a result, all of these methods removed records preferentially within natural subregions with many observations, i.e., proportionate to sampling density especially methods 1 and 4 (Fig. 4). Although method 3 (linear model 2) included natural subregion as model effect, the parameterization of any linear model is still driven by the most common type of data as the algorithm tries to minimize the overall unexplained residual variance across the entire dataset. As a consequence, residual errors of data from sparsely sampled regions are allowed to be larger, and are more likely flagged as outliers.

This is particularly prevalent for method 2 (linear model 1), which does not include natural subregions as a predictor variable. Here, sparsely sampled regions (the Rocky Mountain ecosystems in the southwest and less populated boreal forest areas in the northwest) lose the most records through outlier detection (Fig. 4). These observations are often removed with high confidence (e.g., Fig. 4, large circles for methods 2 and 3 in the Rocky Mountains), but they are not at all flagged by other methods. As a consequence, the residual-based methods 2 and 3 also have the largest effects on regional time series estimates as they remove data points where coverage is already sparse (Table 1). Overall, we consider this attribute of excessive outlier removal in sparsely sampled areas as undesirable. It removes potentially valuable data because it does not fit a model that tries to minimize overall variance. Most notably, this also does not improve Moran's *I* statistics for the overall dataset. As they are spatially remote, the Moran's *I*-based method evaluation correctly allows these observations to be different (Fig. 3).

Our statistical analysis is restricted by a small sample size of only four species that leads to non-significant results. In other word, the variability in how the methods affect different species and their bloom phases is too large, and a species sample size of four is too small to draw general inferences on how the data cleaning methods would work for other species with high confidence. With that caveat in mind, we describe the effects of data cleaning on this database: The method that improved Moran's *I* the most was method 5 (dimensionality reduction and clustering), which is conceptually the most complicated technique, relying on multivariate clustering and dimensionality reduction of a large number of environmental descriptors to provide context to evaluate the consistency phenology observations. This method results in intermediate behavior, with relatively small impacts on regional time series means (Table 1), yet the spatial removal is not focused exclusively on high-density observation areas.

Outliers can generally be classified as point outliers, contextual outliers, or collective outliers (e.g., Aggarwal 2013). The methods that we evaluated have different sensitivities to these outlier types. All of our methods both identify point outliers and use contextual variables. However, the techniques differ in what they use for context. An effective context variable has been the ecoregion delineation used by methods 1 and 3. As an alternative, a high dimensional multivariate approach to provide context proved equally effective (method 5). The only approach that specifically looks for collective outliers—a set of observations from an individual observer—is method 4. Because the latter uses a qualitatively different outlier detection approach, it may be used in complement to one of the other methods.

The removal of inconsistent observers (method 4) also aligns with recommendations by Feldman et al. (2018), based on the observation that trained citizen scientists produced the most precise data. The detection of inconsistent observers is further a means to identify where additional training may be required. We want to emphasize that the use of models to detect inconsistent records should not replace appropriate training, as well as spot checks by trained professionals to assess the overall quality of citizen science data. Nevertheless, model-based data cleaning methods have been widely applied to professional databases, including biodiversity data (e.g., Mathew et al. 2014), genetic data (e.g., Gajer et al. 2004), and ecological databases (e.g., Gueta and Carmel 2016). Similarly, they can and should be used to improve citizen science data.

Conclusions

In summary, both methods 1 and 5 emerged as the best options for outlier removal in citizen science datasets that exhibit spatial sampling biases. It should be noted, however, that the dimensionality reduction (method 5) had very high computational time requirements and is quite complex to execute with several parameter options that need to be optimized. In contrast, method 1 (standardized difference) was computationally and conceptually the simplest approach that could even be carried out with simple spreadsheet-based software. In conclusion, calculating standardized differences for regional data subsets emerges as a simple and effective method to improve reliability of citizen science phenology observations. However, method 1 does require a delineation of ecoregions that also captures differences in sampling density. If such a delineation is not available, method 5 is the best alternative option. Finally, we should note that while method 4 (observer correlation) had the weakest performance overall, the approach is still conceptually sound. Although unable to detect most errors, the method is unlikely to produce false positives in longer time series. Method 4 could therefore be used to remove unreliable observers before applying methods 1 or 5. Lastly, we note that this evaluation of data cleaning methods should be applicable to any large dataset with similar attributes, including datasets compiled by professionally trained research staff.

Acknowledgments We thank all citizen scientists that contributed to the Alberta PlantWatch program led by EB, and we appreciate their enthusiasm and continued support of this program.

Funding information Funding to carry out the analysis presented in this paper was provided by the NSERC Discovery Grant RGPIN-330527 to AH.

References

Aggarwal CC (2013) Outlier analysis. Springer, New York

Beaubien E, Freeland HJ (2000) Spring phenology trends in Alberta, Canada: links to ocean temperature. Int J Biometeorol 44:53–59

Beaubien E, Hamann A (2011a) Spring flowering response to climate change between 1936 and 2006 in Alberta, Canada. Biosci 61: 514–524. https://doi.org/10.1525/bio.2011.61.7.6

Beaubien E, Hamann A (2011b) Plant phenology network of citizen scientists: recommendations from two decades of experience in Canada. Int J Biometeorol 55:833–841. https://doi.org/10.1007/s00484-011-0457-y

Beaubien E, Johnson DL (1994) Flowering plant phenology and weather in Alberta, Canada. Int J Biometeorol 38:23–27

Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59:977–984

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual version 3. www.vsni.co.uk

Crall AW, Newman GJ, Stohlgren TJ, Holfelder KA, Graham J, Waller DM (2011) Assessing citizen science data quality: an invasive species case study. Conserv Lett 4:433–442. https://doi.org/10.1111/j.1755-263X.2011.00196.x

Crall AW, Jarnevich CS, Young NE, Panke BJ, Renz M, Stohlgren TJ (2015) Citizen science contributes to our knowledge of invasive plant distributions. Biol Invasions 17:2415–2427. https://doi.org/10.1007/s10530-015-0885-4

Danielsen F, Jensen PM, Burgess ND, Altamirano R, Alviola PA, Andrianandrasana H, Brashares JS, Burton AC, Coronado I, Corpuz N, Enghoff M, Fjeldså J, Funder M, Holt S, Hübertz H, Jensen AE, Lewis R, Massao J, Mendoza MM, Ngaga Y, Pipper CB, Poulsen MK, Rueda RM, Sam MK, Skielboe T, Sørensen M, Young R (2014) A multicountry assessment of tropical resource monitoring by local communities. Bioscience 64:236–251. https://doi.org/10.1093/biosci/biu001

DataONE (2017) DataONE education module: data quality control and assurance. Data Observation network for Earth. https://www.dataone.org/sites/all/documents/education-modules/pptx/L05_DataQualityControlAssurance.pptx. Accessed 1 Nov 2017

Dickinson JL, Shirk J, Bonter D, Bonney R, Crain RL, Martin J, Phillips T, Purcell K (2012) The current state of citizen science as a tool for ecological research and public engagement. Front Ecol Environ 10: 291–297

Donaldson J (2012) tsne: t-distributed stochastic neighbor embedding for R (t-SNE). R. Package version 0.1–2. http://CRAN.R-project.org/package=tsne

Donnelly A, Yu R (2017) The rise of phenology with climate change: an evaluation of IJB publications. Int J Biometeorol 61(Suppl 1):S29–S50. https://doi.org/10.1007/s00484-017-1371-8

Feldman RE, Zemaite I, Miller-Rushing AJ (2018) How training citizen scientists affects the accuracy and precision of phenological data. Int J Biometeorol 62:1421–1435

Foster-Smith J, Evans SM (2003) The value of marine ecological data collected by volunteers. Biol Conserv 113:199–213

Fraley C, Raftery AE, Murphy B, Scrucca L (2012) mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation technical report no. 597. Department of Statistics, University of Washington

Fuccillo KK, Crimmins TM, de Riviera CE, Elder TS (2014) Assessing accuracy in science-based plant phenology monitoring. Int J Biometerol 59:917–926. https://doi.org/10.1007/s00484-014-0892-7

Gajer P, Schatz M, Salzberg SL (2004) Automated correction of genome sequence errors. Nuc Acids Res 32:562–569

Gueta T, Carmel Y (2016) Quantifying the value of user-level data cleaning for big data: a case study using mammal distribution models. Ecol Informat 34:139–145. https://doi.org/10.1016/j.ecoinf.2016.06.001

Havens K, Vitt P, Masi S (2012) Citizen science on a local scale: the Plants of Concern program. Front Ecol Environ 10:321–323. https://doi.org/10.1890/110258

Hufkens K (2017) khufkens/daymetr: download daymet data using R. Zenodo. https://doi.org/10.5281/zenodo.437886

Hunter J, Alabri A, van Ingen C (2013) Assessing the quality and trustworthiness of citizen science data. Concurrency Computat Pract Exper 25:454–466. https://doi.org/10.1002/cpe.2923

IPCC (2007) Intergovernmental Panel on Climate Change, Climate Change 2007: synthesis report. Contribution of Working Groups I, II, and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Core Writing Team, Pachauri RK, Reisinger A (eds). IPCC, Geneva, Switzerland, 104 pp

Kosmala M, Wiggins A, Swanson A, Simmons B (2016) Assessing data quality in citizen science. Front Ecol Environ 14:551–560. https://doi.org/10.1002/fee.1436

MacKenzie CM, Murray G, Primack R, Weihrauch D (2017) Lessons from citizen science: assessing volunteer-collected plant phenology data with Mountain watch. Biol Conserv 208:121–126. https://doi.org/10.1016/j.biocon.2016.07.027

Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams A, de Jong Y, Goble C (2014) A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. Biodivers Data J 2:e4221. https://doi.org/10.3897/BDJ.2.e4221

McKinley DC, Miller-Rushing AJ, Ballard HL et al (2017) Citizen science can improve conservation science, natural resource management, and environmental protection. Biol Conserv 208:15–28. https://doi.org/10.1016/j.biocon.2016.05.015

Mehdipoor H, Zurita-Milla R, Rosemartin A, Gerst KL, Weltzin JF (2015) Developing a workflow to identify inconsistencies in volunteered geographic information: a phenological case study Plos One 10. https://doi.org/10.1371/journal.pone.0140811

Miller-Rushing A, Primack R, Bonney R (2012) The history of public participation in ecological research. Front Ecol Environ 10:285–290. https://doi.org/10.1890/1102798

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika. 37(1):17–23

Natural Regions Committee (2006) Natural regions and subregions of Alberta. Compiled by D.J. Downing and W.W. Pettapiece. Edmonton. Pub. No. T/852. Alberta Environment, Government of Alberta, Edmonton, AB

Paradis E, Claude J, Strimmer K (2004) APE: analysis of phylogenetics and evolution in R language. Bioinformatics 20:289–290

R Development Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna http://www.R-project.org/

Ranjitkar S (2013) Effect of elevation and latitude on spring phenology of rhododendron and Kanchenjunga conservation area, East Nepal. Int J Appl Sci Biotech 1:253–257. https://doi.org/10.3126/ijasbt.v1i4.9154

Rathcke B, Lacey EP (1985) Phenological patterns of terrestrial plants. Ann Rev Ecol Syst 16:179–214

Schwartz MD, Beaubien EG, Crimmins TM, Weltzin JF (2013) Chapter 5. North America. In: Schwartz M (ed) Phenology: an integrative environmental science. Springer, Dortrecht, pp 67–89

Schwartz MD, Hanes JM, Liang L (2014) Separating temperature from other factors in phenological measurements. Int J Biometeorol 58: 1699–1704. https://doi.org/10.1007/s00484-013-0723-2

Silvertown J (2009) A new dawn for citizen science. Trends Ecol Evol 24:467–471

Silvertown J, Buesching CD, Jacobson SK, Rebelo T (2013) Citizen science and nature conservation. In: Macdonald DW, Willis KJ (eds) Key topics in conservation biology 2, 1st edn. Wiley, New York, pp 127–142

Thornton PE, Thornton MM, Mayer BW, Wilhelmi N, Wei Y, Devarakonda R, Cook RB (2016) Daymet: daily surface weather data on a 1-km grid for North America, Version 3 ORNL DAAC, Oak Ridge, Tennessee, USA. Accessed June 5, 2017. Time period: 1987-01-01 to 2016-12-31. Spatial range: N=59.82, S=49.13, E=-109.22, W=-119.67. https://doi.org/10.3334/ORNLDAAC/1219

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234–240. https://doi.org/10.2307/143141

USANPN (2017) USA National Phenology Network. How to observe. https://www.usanpn.org/nn/guidelines. Accessed 2 Nov 2017

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:1–48

Vander Stelt E, Fant JB, Masi S, Larkin DJ (2017) Assessing habitat requirements and genetic status of a rare ephemeral wetland plant species, *Isoëtes butleri* Engelm. Aquat Bot 138:74–81. https://doi.org/10.1016/j.aquabot.2017.01.002

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.