VGI users & data centered methods for the analysis of farmland biodiversity indicators open issues

Sandro Bimonte
IRSTEA TSCF
9 Avenue Blaise Pascal
63178 Aubiere, France
sandro.bimonte@irstea.fr

Aurélien Besnard
LPO Aquitaine
433 Chemin de Leysotte
33140 Villenave-d'Ornon, France
aurelien.besnard@lpo.fr

Elodie Edoh-Alove,
GEOSYSTEMS France
6 rue Jean Pierre Timbaud
78180 Montigny-Le-Bretonneux, France
edoh-alove@geosystems.fr

Ali Hassan
CESCO, Muséum National d'Histoire naturelle
43 Rue Buffon
75100 Paris, France
ali.hassan@mnhn.fr

Karine Princé
CESCO, Muséum
National d'Histoire naturelle
43 Rue Buffon
75100 Paris, France
karine.prince@mnhn.fr

Amir Sakka
IRSTEA, TSCF
9 Avenue Blaise Pascal
63178 Aubiere, France
amir.sakka@irstea.fr

Pascale Zarate
IRIT, Université Capitole
2 Rue du Doyen-Gabriel-Marty
31042Toulouse, France
Pascale.Zarate@ut-capitole.fr

Abstract

Motivated by the importance of the analysis of farmland biodiversity data, and the lack of advanced analysis tools of VGI systems, in this paper we present main issues related to the analysis of VGI farmland biodiversity data using SOLAP systems. We develop challenges related to volunteers and crowd sourced data. Then, we present some possible solutions that represent the main work of our French ANR project VGI4Bio

Keywords: Farmland biodiversity Indicators, Volunteered Geographic Information and Community Observatories, Spatial OLAP, Geospatial Technologies for All

1. Introduction and motivation

The monitoring and conservation of biodiversity in farmlands currently represent major challenges, agriculture being the dominant land use in Europe and biodiversity in these landscapes ongoing rapid and massive decline due to intensive agricultural practices (Bommarco et al., 2013). Besides, many promising alternatives to improve the sustainability of agriculture rely on the ecosystem services provided by biodiversity (Prince et al., 2012). However, financial and human resources may be limited to collect the data needed to measure impacts, assess effectiveness of conservation policies or changes in agricultural practices and forecast future changes. To build the biodiversity indices used in these assessments, observation data are needed at large spatial and temporal scales to encompass a wide range of situations, and are usually provided through standardized monitoring schemes. Large numbers of observers need to be mobilized, at a cost which would be prohibitive (Reginer et al., 2015) unless they are volunteers in citizen science programs. However, a wealth of data on biodiversity outside any standardized framework is produced in the course of leisure activities. These data, collected at almost no financial cost at large spatial and temporal scales are currently poorly exploited, because of statistical challenges In this context, VGI technology (Volunteered Geographic Information), defined by (Sui et al., 2013) as "the mobilization of tools to create, assemble and disseminate geographic data provided by volunteers" allows to manage large amount of geolocalized data and is widely used in different application domains. Therefore, we suggest that the use of VGI technology in participative monitoring of biodiversity would have important social, economic and environmental benefits.

However, VGI systems do not support advance analysis tools of GeoBusiness Intelligence (GeoBI) systems. GeoBI systems allow stakeholders to analyze geo-referenced indicators using cartographic displays (Golfarelli et al., 2013). We argue that GeoBI technologies, and in particular Spatial Data Warehouse (SDW) and Spatial OLAP (SOLAP) can be successfully used to analyze VGI data, and should be developed for farmland biodiversity monitoring. A SDW is “a collection of subject-oriented, integrated, non-volatile and time-variant spatial and non-spatial data to support the decision-making process” (Bédard et al., 2007). Warehoused spatial data are modeled according to the spatio-temporal multidimensional model, which defines the concepts of spatial dimensions (analysis axes) and spatial measures (analysis subjects). This multidimensional data structure allows the online analysis provided by SOLAP systems. SOLAP systems are "visual platforms built especially to support rapid and easy

spatiotemporal analysis and exploration of data, following a multidimensional approach, comprised of aggregation levels, available in cartographic displays as well as in tabular and diagram displays" (Bédard et al., 2007). Since SDWs are conceived according to data sources and users requirements, the more the SDW model reflects stakeholders' needs, the more the stakeholders will make use of their data, implying social (e.g. welfare improvement) and economical (e.g. sustainable agriculture) benefits. Therefore, providing volunteers with GeoBI applications fitting their particular needs represents important social and economic advances.

In this context, we present the main challenges of taking into account the particularities of VGI data and users for the definition of a SOLAP system developed to analyze farmland biodiversity.

The paper is structured as following: Section 2 presents the main challenges related to VGI farmland biodiversity users, Section 3 highlights open issues of VGI farmland biodiversity data, and finally Section 4 introduces our French ANR project VGI4Bio, which aims at addressing these issues.

2. VGI users open issues

In this section, we define open issues related to taking into account the diversity of volunteers in the analysis of biodiversity data.

Challenge I: Participative design of SOLAP models

(S)DWs design has been investigated in several works (Romero et al., 2009). Three types of approaches have been defined: (i) methods based on user specification (user-driven approach), which define the DW schema using users requirements only (i.e. analysis needs); (ii) methods based on data sources (data-driven approach), where the multidimensional schema is automatically derived from the data sources; (iii) mixed methods (mixed approach), which merge data-driven and user-driven methodologies. Analysis needs within user-driven approaches are formalized using complex formalisms such as UML and ER and/or using declarative query languages (i.e. SQL) (Romero et al., 2009)). Although several systems allow collaborative conceptual design for generic applications (Wang et al., 2002), and more recently for collaborative GIS (Roche et al., 2012), existing DWs design methodologies are not implemented in such kind of tools, since they are not designed for multi-users. These approaches only focus on the translation of conceptual requirement models into the multidimensional schema, without detailing how users create them. Only (Corr et al., 2011) provide an agile questionnaire-based methodology to help decision-makers to work together in the conception of the DWs, but this approach does not consider decision-makers one by one with their preferences, and it is not supported by a computer tool. Therefore, it cannot be considered as a participative design of multidimensional databases.

Challenge II: Rapid prototyping of SOLAP models

Usually, formalizing users needs allows for a rapid prototyping methodology. Indeed, some attempts have been made to apply agile practices to DW design (Corr et al., 2011) . The main methodological principles used to this end are incrementally and iteration, prototyping, user involvement and automated schema transformation. An example of this kind of methodologies is ProtOLAP methodology, implemented in a relational architecture (Bimonte et al., 2013). However, these methodologies do not take into account geovisualization analysis needs of decision-makers, which make not effective alphanumeric DW prototyping methodologies. Indeed, it has been widely recognized that the SOLAP decision-making process is based on effective cartographic representations. Geovisualization methods that do not fit with cartographic mind representations of decision-makers are not suitable for a successfully SOLAP project.

3. VGI data open issues

In this section, we define issues related to data quality for biodiversity monitoring.

Challenge III: Quantity vs. Quality

Biodiversity indicators may be specific and report trends in relative abundance of given species, which is the first aim of biodiversity monitoring schemes and can have value for particular users (e.g. farmers monitoring a given pest species). Alternatively, monitoring the frequency of species sharing ecological, biological or other traits allows to document phenomena such as community homogenization (decline of farmland or woodland specialists) or response to climate change (increase of warm-adapted species) (Devictor et al., 2008). These composite indicators are widely used by researchers, but also by managers and policy-makers: the Farmland Bird Index, based on common birds all over Europe, is part of the EU Sustainable Development Goals (SDG) indicator set. Whatever the indicator, interest will often not be in abundance itself but in its variation in space and time. Statistically valid sampling designs and standardized protocols are recommended for collecting the data. However these protocols are constraining, and participation may not be enough to produce sufficient high-quality data to get meaningful indicators. Standardized data rarely have sufficient spatial and temporal coverage. On the other hand, opportunistic data, of lower quality because produced without standardization, are collected routinely by thousands of nature-lovers, and stored in databases such as Biolovision in several European countries. These data, despite being very abundant, cannot be used with current statistical tools because of data quality issues and the difficulty to model observer behaviour outside standardized schemes

Challenge IV: Complex SOLAP models for biodiversity analysis

Integration of VGI data into SOLAP systems has been investigated only by (Bimonte et al., 2014). Using a real-world scenario, authors highlight similarities and differences among these systems and define a conceptual quality-oriented framework for warehousing and OLAPing VGI data. In particular, to address precision and credibility problems related to VGI data, they propose two new ETL operators: aggregation based on the VGI credibility and a filter based on historical precision. There are few publications using (S)OLAP technologies in the environmental domain ((Bimonte, 2016) for a survey). Finally, (Sautot, et al., 2015) present biodiversity model using (S)OLAP technologies to

study bird population. They propose a SDW model to address issues such as "What is the total abundance per year and census point?" However, these works do not take into account the quality of VGI data as previously described in the Challenge *Quantity vs Quality*.

4. Some possible solutions: the VGI4Bio project

The issues described above will be addressed in the French ANR project VGI4Bio (www.VGI4Bio.fr). VGI4Bio is started on 4th December 2017 and it will finish on December 2021. Partners involved in the project are: IRIT, Irstea, CESCO, LPO Aquitaine and GEOSYSTEMS France. In this project, we have identified some interesting research lines to solve the above described issues.

Challenge I: Participative design of SOLAP models

To address this issue, we suggest an innovative SDW design methodology based on participative Group Decision-making Support System (GDSS). GDSS are designed to support group engaged in a collective decision process (Zaraté, 2013). Intended to provide computational support to participative decision-making processes, GDSS represent a widely used collaborative technology which increases user participation and decision-making quality. The GRUS system (Zaraté, 2013), developed at IRIT, offers the basic services commonly available in GDSS and Collaborative Systems. Participative work allows users to exchange, produce, share and modify information and knowledge without physical or temporal barrier. These methodologies are used in several domains such as workflows, user interface and databases (Wang et al., 2002), but not in SOLAP context. At the moment, we are defining SOLAP models with volunteers and we will aggregate them with our new participatory methodology.

Challenge II: Rapid prototyping of SOLAP models

To solve this challenge, we suggest to define a geovisualization model for SOLAP and integrate it to existing OLAP prototyping methodologies and tools (Bimonte et al., 2016). Geovisualization analysis needs must be expressed at a conceptual level and then automatically implemented. Therefore, we propose to extend existing conceptual models for SDW with geovisualization elements, and provide their automatically representation. A preliminary work is an extension of the DW prototyping methodology (Bimonte et al., 2016) with cartographic elements.

Moreover, SOLAP maps are usually defined by hand. Decision-makers spend time to obtain a readable map for each SOLAP query. It delays the decision-making process, and it avoids on-line exploration of spatial warehoused data. Therefore, we propose to introduce some intelligent algorithms that automatically choice for the best readable cartographic representation of each SOLAP query. A first work that addresses the number of displayed graphic elements of a map to grant a readable visual representation has been proposed in (Bimonte et al., 2016) . We will extend this work to include other parameters related to the readability of SOLAP maps: spatial object features (i.e. number of points, distance among objects), visual cluttering of thematic maps, etc.

Challenge III: Data quality

A promising way to overcome data restrictions and data quality issues is the combination of different data types from various sources that contain information on the occurrence and abundance of a species across space and time. A new field of research in statistical ecology has recently emerged, and with it the development of new more sophisticated analytical approaches (Calenge et al., 2015), allowing the analysis of opportunistic data, numerous and collected at almost no cost (except website and database conception and maintenance). As an example, where citizen science data are subject to random sampling errors, mixed-effects models have proven extremely useful in ecological studies. However, systematic bias must be dealt with using other approaches, such as hierarchical models, which were created to account for detection bias (MacKeinze et al., 2005), but have potential to handle unknown and varying observation efforts. This type of hierarchical 'state-space' model makes it possible to explicitly model the latent state-variables of interest (i.e. occupancy, colonization, extinction) as distinct from the observation process (i.e. detection) yielding the observed data. In addition, combining opportunistic data with data collected through schemes characterized by a known sampling effort results in more accurate estimations of temporal trends than with standardized data alone, particularly for rare species (Giraud et al., 2015). Another independent, parallel approach on the use of opportunistic data has also been recently published (Fithian et al., 2015). However, several improvements of these methods are needed. For instance, spatial heterogeneity of the surveyed area should be taken into account, especially the possibility that observational biases toward some habitat types may vary across different sites. Another improvement would be taking into account spatial and temporal autocorrelation, i.e. the increase in similarity in species densities as sites are closer or habitats or dates are more similar. Last but not least, the approach of (Giraud et al., 2015) assumes that measurement errors are negligible, which may be wrong since even experts may misidentify species, and even a few false positive can bias estimates. This issue could be overcome by combining data coming from different sources, with different error rates, but this should be tested (Millet et al., 2011).

Challenge IV: Complex SOLAP models for biodiversity analysis

In order to integrate farmland biodiversity data into SOLAP models, we propose to develop some constellation models (i.e. models composed of more facts) with VGI data and indicators obtained using statistical methods above described.

5. Conclusion

Motivated by the importance of the analysis of farmland biodiversity data, and the lack of advanced analysis tools of VGI systems, in this paper we present main issues related to the analysis of VGI farmland biodiversity data using SOLAP systems. We develop challenges related to volunteers and crowd sourced data. Then, we present some possible solutions that represent the main work of our French ANR project

VGI4Bio. The project will create a new interface creating bonds between citizens and the scientific world. More than data access and visualization, it will give the opportunity to explore and play with data collected through citizen science. This will develop synergies between researchers and citizens, may raise new questions and new results. Having access to biodiversity data and statistically sound indicators will help reconnection to nature by making anthropic impacts and biodiversity-rich areas visible, and understanding that the surrounding environment is infinitely more complex than what is usually imagined. Access to information on programs, and to data and results coming from participative science dealing with anthropic impacts on biodiversity is an important step forward to allow anybody to take informed decision to act and get involved, and hence to reduce inequalities and favor positive citizenship.

Acknowledgement

This work is supported by French Agence National de la Recherche for the ANR project ANR-17-CE04-0012

References

Bedard, Y., Rivest, S. and Proulx, M.-J (2007). Spatial Online Analytical Processing (SOLAP): Concepts, Architectures, and Solutions from a Geomatics Engineering Perspective. *Data Warehouses and OLAP: Concepts, Architectures and Solutions*, pp. 298–319.

Bimonte, S., Boucelma, O., Machabert, O. and Sellami, S. (2014) A new Spatial OLAP approach for the analysis of Volunteered Geographic Information. *Computers, Environment and Urban Systems* 48, 111-123.

Bimonte, S., Edoh-Alove, E., Nazih, H., Kang, M. and Rizzi, S. (2013) ProtOLAP: rapid OLAP prototyping with on-demand data supply. In: *Proceedings of* DOLAP 2013, 2013, pp. 61-66.

Bimonte, S. (2016). Current approaches, challenges and perspectives on Spatial OLAP for Agri-Environmental Analysis. *Journal of Agricultural and Environmental Information Systems*, 7 (4), 33-50.

Bimonte, S, Hassan, A. and Beaune, P. (2016). From Design to Visualization of Spatial OLAP Applications: A First Prototyping Methodology. In: *Proceedings of ER Workshops 2016*, 2016, pp. 113-123.

Bommarco R, Kleijn D & Potts SG. 2013. Ecological intensification: harnessing ecosystem services for food security. Trends in ecology and evolution. 28: 230-238.

Calenge, C., Chadoeuf, J., Giraud, C., Huet, S., Julliard, R., Monestiez, P., Piffady, J., Pinaud, D. and Ruette, S. (2015). The Spatial Distribution of Mustelidae in France. *Plos One*, 10.

Connors, J. P., Lei, S. and Kelly, M. (2012). Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers*, 102(6), 1267-1289.

Corr, L. and Stagnitto, J. (2011). *Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema*. DecisionOne Press.

Devictor, V., Julliard, R. and Jiguet, F. (2008). Distribution of specialist and generalist species along spatial gradients of habitat disturbance and fragmentation. *Oikos* 117, 507–514.

Fithian, W., Elith, J., Hastie, T. and Keith, D.A (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol. Evol.* 6, 424–438.

Giraud, C., Calenge, C., Coron, C. and Julliard, R. (2015). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, 72, 649–658.

Golfarelli, M., Mantovani, M. and Ravaldi, F. (2013). Lily: A Geo-Enhanced Library for Location Intelligence. In: *Proceedings of DaWaK 2013*, pp. 72-83.

Wang, L., Shen, W., Xie, H., Neelamkavil, J. and Pardasani, J. (2002). Collaborative conceptual design - state of the art and future trends. *Computer-Aided Design* 34(13), 981-996

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L., Hines, J.E. 2005. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. and Weir, L.A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology* 92, 1422–1428.

Prince, K., Moussus, JP. and Jiguet, F. (2012). Mixed effectiveness of French agri-environment schemes for nationwide farmland bird conservation. *Agriculture Ecosystems & Environment*,149, 74-79

Regnier, C., Achaz, G., Lambert, A., Cowie, R.H., Bouchet, P. and Fontaine, B. (2015). Mass extinction in poorly known taxa. *Natl. Acad. Sci. U. S. A.* 112, 7761–7766.

Roche, S., Mericskay, B., Batita, W., Bach, M. and Rondeau, M. (2012). WikiGIS Basic Concepts: Web 2.0 for Geospatial Collaboration. *Future Internet*, 4(1), 265-284

Romero, O. and Abelló, A. (2009). A Survey of Multidimensional Modeling Methodologies. *Int. Journal of Data Warehousing and Mining*, 5(2), 1-23.

Sautot, L., Faivre, B., Journaux, L. and Molin, P. (2015). The hierarchical agglomerative clustering with Gower index: A

methodology for automatic design of OLAP cube in ecological data processing context. Ecological Informatics, 26(2), 217-230.

Sui, D.Z., Elwood, S. and Goodchild, M. (2013). *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer.

Zaraté. P. (2013). *Tools for Collaborative Decision-Making*, Wiley.