Quality Indicators of a Wetlands Data base in Canada: An Environmental Data Analysis Case Study

Herbert Nkwimi Tchahou[1], Claude Girard [2] and Martin Hamel[3]

ABSTRACT

In order to monitor wetlands, Environment Canada (EC) has access to mega-databases containing a large quantity of information describing the many facets of the Canadian territory. A pilot project implemented jointly by EC and Statistics Canada looked into assessing the quality of these databases – which possess attributes of Big Data, administrative data and survey data – using finite population survey and data analysis techniques. In this paper, we give an overview of the methodology used as part of a model validation exercise.

KEY WORDS: Data Analysis; Environment; Quality Assessment.

RÉSUMÉ

Afin d'assurer le contrôle des terres humides, Environnement Canada (EC) dispose de méga-bases de données contenant une masse importante d'informations diverses décrivant le territoire canadien sous toutes ses facettes. Dans le but de juger de la qualité de ces bases - qui présentent à la fois certains des attributs propres aux données volumineuses (« Big Data »), aux données administratives et aux données d'enquêtes - nous avons exploité des techniques de sondage de populations finies et d'analyse de données. Dans cet article, nous présenterons un projet pilote, mené conjointement par Statistique Canada et Environnement Canada : un aperçu de la méthodologie employée dans l'exercice d'un modèle de validation.

MOTS CLÉS : Analyse de données; Environnement; Évaluation de Qualité.

1. INTRODUCTION

1.1 Description of the Problem

Wetlands, which include ecosystems with water-saturated soil, play an important role in the survival of many species. In order to facilitate their preservation and long term management there is a classification system designed to gather and organize wetland information. Based on the geophysical literature, experts generally agree that there are five main types of wetlands: shallow water (I), marshes (II), swamps (III), ombrotrophic bogs (IV) and minerotrophic bogs (fens) (V). In Canada, information about wetlands is gathered in various databases (one for each province and territory) in a detailed and hierarchical manner. Record entities or units in these databases are geographic parcels called wetland polygons. These databases also contain various cartographical tools and products used to locate wetlands. Information contained in these databases can be used to classify each polygon into one of the above predefined categories. One way to proceed is to have experts classify the polygons one by one which may require on-site visits. While appealing, this option cannot be carried out in practice given the amount of time and the resources it would require. Indeed, the smallest database contains thousands of polygons to be classified (see Nkwimi et al 2014). Another option is to classify polygons using a model derived by subject-matter experts. The main advantage of this approach is the gain in time and resources it brings. As with any modeling exercise, the classification produced by this model may not concur with what experts would agree on; thus, mis-classification is to be expected. This is why some form of validation is required to evaluate the modeling approach. To help validate the classification put forward by the model one can select a small random sample of polygons and submit those to experts. How much disagreement is there in the sample between expert and model and what does it say about the

1 ,3: Business Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON, Country Canada, K1A 0T6, Herbert.NkwimiTchahou@canada.ca, Martin.Hamel@canada.ca
2: Statistical Research And Innovation Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON, Country Canada, K1A 0T6, Claude.Girard @canada.ca

automated classification produced at the scale of the whole database? These are some of questions we try to answer through a validation process. This paper is the second of a two-part paper. The first part, Nkwimi et al. (2014), analyzed and structured the information contained in these databases, paving the way for a statistical validation analysis to be performed. In this paper, we explore the methodology framework used to carry out the validation process of the automated classification mentioned above. In the following we assume that we have a database of polygons with two types of classifications: one from a model, available for the entire database and a second one, from experts, which is available only for a small stratified random sample of polygons. By using the classification resulting from the model as a stratification variable we were assured that all categories used by the model were represented in the sample. In order to validate the model's classification, two types of analyses were conducted: first, we derived a macro quality indicator using finite population survey techniques and, second, we used data mining techniques to get a micro quality indicator. Whereas the macro indicator is a statement about the level of agreement for the entire database, the micro indicator speaks of the concordance between the two classifications polygon by polygon. The rest of the paper is organized as follows: the macro and micro indicators are presented in Section 2 and Section 3 respectively, and a brief conclusion is given in Section 4.

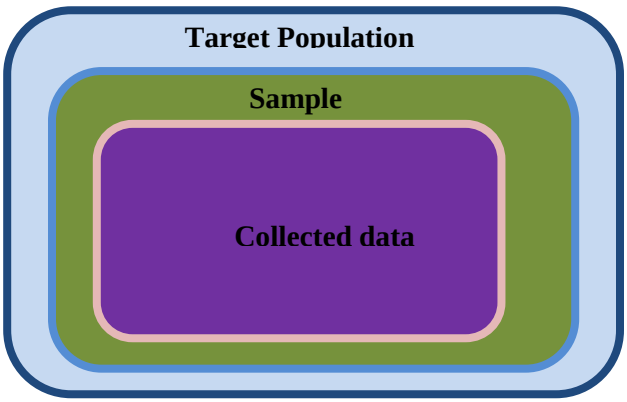
2. Analysis I: Macro Quality Indicator

In this section, we derive a macro quality indicator for the database using finite population survey techniques. More precisely, we estimate based on the findings from a small sample the number of times both expert and model classifications would have agreed had the expert been able to examine all polygons in the database.

2.1 Framework and Methodology

To see how the situation here fits within regular traditional surveys, we draw the following parallels. The target population is the entire data base of polygons and the unit of interest is the polygon. The sample is the set of polygons for which an expert classification is available. Thus, we are in the unusual survey situation where the response set matches the whole sample: the expert by assumption did classify every single one of the sampled polygons. The sample design is a stratified simple random sample and the auxiliary variable used for stratification is the classification produced by the model. One can see Särndal et al. (1992), for more details on survey sampling. The following figure summarizes a typical framework.

Figure 1 – framework for a typical



In our case, the two inner rectangles coincide. In terms of estimation, each polygon i gets a design weight d_i which is the inverse of its inclusion probability. These weights are used to inflate reported values in the Horvitz-Thompson estimator. Since the response rate is 100%, no weight adjustment is needed. For a given category of polygon C, the number of polygons of type C and proportion of polygons of type C are estimated as follow:

$$\widehat{N}_C = \sum_{i \in S \cap C} d_i \text{ : Number of polygons of type } C$$

$$\hat{P}_C = \frac{\sum_{i \in S \cap C} d_i}{\sum_{i \in S} d_i} \text{ : Proportion of polygons of type } C,$$

where *s* is the stratified sample of polygons.

2.2 Interpretation of possible outcomes

The estimates described in the previous section can be used to measure disagreement between the automated classification and the expert at the scale of the whole database. For a given type of wetlands, we will either find that the estimate of the number of polygons closely matches the count yielded by the model or significantly differs from it. If the latter case happens, then we can conclusively say that there is a strong disagreement between model and expert on how many wetlands of that type are presumed to exist in the database. On the other hand, should the estimate and model counts be a near match we would need to be very cautious in interpreting this as a strong agreement between expert and model. Indeed, it is not because both would agree that *p*% of all polygons are of a given type that they agree on *which* polygons are of that type. In other words, do the polygons identified by the model as being of a certain type match exactly the ones identified by the expert? We need a more detailed analysis to answer this question. This is the subject of the next section.

3. Analysis II: Micro Quality Indicator

In this section, we derive a quality indicator on a polygon basis. More precisely, based on sampled polygons and the consultant's classifications, we build a rule to assign each polygon of the entire data base to one of the five categories. Two types of analyses were explored: in a descriptive approach, we used the conditional Cohen Kappa coefficient to assess the degree of concordance between model and consultant and, in a second approach; we built a generalized logistic regression model to create a predictive rule to classify the polygons.

3.1 Conditional Cohen Kappa Coefficient

When two or more observers independently classify a set of *n* items in the same set of *I* mutually exclusive categories, it is often of interest to measure the degree of agreement between them. The Kappa coefficient is one of the most widely used methods to assess such agreement, especially in social sciences. The initial version of the coefficient was introduced by Cohen (1960). Intuitively, it provides a composite measure of agreement across all categories and observers. Cohen's seminal paper has since been generalized in many directions. For instance, there is the weighted version proposed by Cohen (1968) and the conditional agreement measure introduced by Coleman (1966), just to name a few. A weighted version is useful in situations where some disagreements are more important than others. With the Coleman measure of conditional agreement, one of the observers acts as a gold standard: the verdict of all other observers is assessed on items that this one observer has already classified as type *i*. Since we are validating the classification induced by the model and thus can naturally see the verdict of the expert as the gold standard, the conditional agreement measure is more convenient for our purposes here. To illustrate, consider the following fictitious numerical example opposing the verdicts of two observers, model and expert, using a $\mathbf{5 \times 5}$ Contingency table:

Table 2 – Contingency table: Model vs. Expert

		Modeling					
	Polygon type	**I**	**II**	**III**	**IV**	**V**	**Total**
EXPERT	**I**	65	5	20	10	0	100
	II	10	75	50	10	5	150
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Total	300	200	150	225	125	1,000

According to fictitious data reported in Table 2, the expert has identified 100 polygons as of type I. Of these, how many actually were identified as type I polygons by the model? The conditional Kappa coefficient can be computed as follows (see Tarald (1985) for more details):

$$\tau = \frac{P_{ii} - P_{i+}P_{+i}}{P_{i+} - P_{i+}P_{+i}}$$

P_{ii}, the diagonal term, represents the probability that both observers classified an item into the same category i. For example based on Table 2, $P_{11} = 65/1{,}000 \sim 0.70\%$; $P_{i+} = \sum_{j=1}^{I} P_{ij}$ is the marginal total of row i; and $P_{+j} = \sum_{i=1}^{I} P_{ij}$ is the marginal total of column j.

3.2 Outcomes of Conditional Agreement Measure

The Kappa coefficient is similar to a correlation coefficient between two variables and is generally used as a descriptive statistic rather than as part of a formal statistical hypothesis test of agreement. It is important to note that chances to obtain a low Kappa are generally greater when dealing with a high number of categories. This reflects the fact that two evaluations of the true colour of items, for instance, are more likely to agree on a matter of black versus white than if various shades of grey are introduced. Therefore, describing the strength of agreement associate with a given Kappa coefficient remains a challenge. There is no universal guideline on how to interpret the Kappa in the literature. One of the interpretation tables commonly used and presented below (this is what was used in this project) is provided by Landis and Koch (1977).

Table 3 –Strength of agreement base on Kappa

Kappa	Strength of Agreement
<0.00	Poor
0.00-020	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1	Almost perfect

The conditional Kappa may give insights into the issues raised earlier about Analysis I. For example, in the case where automated classification and expert both agree that *p*% of all polygons are of a given type, a conditional Kappa value below Substantial would be indication that they do not agree as to *which* ones actually are of that type. Thus, with the conditional Kappa, one may be able to identify categories of polygons more likely to be involved in misclassification.

So far, we have explored two different tools (finite population sampling and Kappa coefficient) in order to assess the agreement between classifications from an automated modeling algorithm and expert. However, we are not able to tell based on these analyses how susceptible any given polygon is to be misclassified. To be able to provide a verdict on a polygon basis, we have performed a generalized logistic regression to build a classification rule. More details are given in the following section.

3.3 Generalized Logistic Regression

In this section, we discuss a generalized logistic regression in order to identify potentially misclassified polygons. The expert and model classifications, which again are available for the sampled polygons, are used to build a rule to classify each polygon into a particular category. The predictive performance of our model (success rate) can be evaluated by the total fraction of the sampled polygons that are correctly classified by the model. One can see Pearce and Ferrier (2000) for more details about predictive performance. It is common to split the sample into two groups: one learning sample used to build the rule, and a test sample used to evaluate that rule. However, given the relatively small size of the initial sample, we used a replicate method rather than split the sample. We fit two different models. In a first basic model, we predicted the expert classification based on modeling classification only. The rule built from that model was used to classify sampled polygons. We found that, the rule from this basic model performs very well in identifying some types of polygons (very high success rate). The global success rate was also relatively high given the simplicity of the model.

In order to try to improve this global rate, we created a second model by adding new independent variables to the basic model. These new variables were chosen from the set of available variables in the dataset. A backward selection process was used to select the variables. In the end, we added two other new variables in the model in addition of modeling classification. The rule built from this enriched model resulted in a global success rate which appears to be only marginally higher than that from the basic model. We think that a possible reason why the augmented model did not produce a significantly higher global success rate is that the information from the extra variables is already a factor in the modeling automated classification process. However, even though the increase in percentage is small, it represents a significant number of polygons overall.

4. Conclusion

This pilot project looked into the performance of an automated classification algorithm of wetlands. We started with a mass of data that was very rich but non-structured, making it difficult to use. The data presented attributes reminescent of Big Data, administrative data and survey data. We first organized, analyzed and summarized the available information and then we performed a validation of an automated classification algorithm for polygons. In the end, we derived two types of quality indicators, one macro and the other micro. The study could be pursued further by exploring other methods. For instance, in order to improve the global success rate of the rule built from generalise logistic regression, one could fit a separate model and different classification rule for each type of polygon instead of a global model as done here. Another possible exploration avenue could be to perform logistic regression on independent factors retrieved from a factorial analysis.

5. Acknowledgements

We would like to thank Jeannine Morabito, Christian Olivier Nambeu, Assoumou Ndong Franklin, Nathalie Hamel and Wesley Yung for all their efforts in reviewing this paper.

REFERENCES

Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational & Psychological Measurement*, **20**, 37-46.

Cohen, J. (1968). "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". *Psychological Bulletin*, **70**, 213-220.

Coleman, J.S. (1966). "Measuring Concordance in Attitude". Unpublished manuscript, *Johns Hopkins University*, Department of Social Relations, Baltimore.

Landis, J. R., and Koch, G. G. (1977). "The measurement of observer agreement for categorical data". *Biometrics* **33** (1):159-174.

Nkwimi, T.H., Girard, C., Hamel, M. (2014). "Making Use of Administrative, Big and Survey Data: An Assessment of the Quality of Canadian Wetland Databases". *Proceedings of Statistics Canada Symposium 2014*.

Pearce, J., Ferrier, S. (2000). "Evaluating the Predictive Performance of Habitat Models developed using Logistic Regression". *Ecological Modelling* **133** (2000) 225–245.

Särndal, C.E., Swensson, B. and Wretman, J.(1992). *Model Assisted Survey Sampling*. Springer-Verlag, NY.

Tarald, O. K. (1985). "Weighted conditional kappa". *Bulletin of the Psychonomic Society* 1985, **23** (6), 503-505.