Remote Sensing of Environment 176 (2016) 31–42

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Triple collocation for binary and categorical variables: Application to validating landscape freeze/thaw retrievals

Kaighin A. McColl [a,*], Alexandre Roy [c], Chris Derksen [d], Alexandra G. Konings [a], Seyed Hamed Alemohammed [a], Dara Entekhabi [a,b]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Centre d'Application et de Recherches en Télédétection, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[d] Environment Canada, Toronto, ON M3H 5T4, Canada

ARTICLE INFO

Article history:
Received 3 September 2015
Received in revised form 11 January 2016
Accepted 16 January 2016
Available online xxxx

Keywords:
Triple collocation
Freeze/thaw classification
SMAP
Aquarius

ABSTRACT

Triple collocation (TC) can be used to validate observations of a continuous geophysical target variable when the error-free true value is not known. However, as we show in this study, naïve application of TC to categorical target variables results in biased error estimates. The bias occurs because the categorical variable is usually bounded, introducing correlations between the errors and the truth, violating TC's assumptions. We introduce Categorical Triple Collocation (CTC), a variant of TC that relaxes these assumptions and may be applied to categorical target variables. The method estimates the rankings of the three measurement systems for each category with respect to their balanced accuracies (a binary-variable performance metric). As an example application, we estimate performance rankings of landscape freeze/thaw (FT) observations derived from model soil temperatures, in-situ station air temperatures and satellite-observed microwave brightness temperatures in Alberta and Saskatchewan, Canada. While rankings vary spatially, in most locations the model-based FT product is ranked the highest, followed by the satellite product and the in-situ air temperature product. These rankings are likely due to a combination of differences in measurement errors between FT products, and differences in scale. They illustrate the value in using a suite of different measurements as part of satellite FT validation, rather than simply treating in-situ measurements as an error-free 'truth'.

1. Introduction

Categorical variables belong to one of a set of exhaustive, mutually-exclusive categories, which may be ordered (in which case, the categorical variable is 'ordinal') or unordered ('nominal'). For many geophysical variables, it is convenient to consider the variable to be categorical rather than continuous. Examples include land cover type (Friedl et al., 2002), cloud presence/absence (Ackerman et al., 1998), wildfire burned area status (Roy, Boschetti, Justice, & Ju, 2008) and landslide occurrence (Metternicht, Hurni, & Gogu, 2005). Models, satellites and in-situ observations (or "measurement systems") are used to monitor and understand these variables, but each system contains its own errors. A common question is: which system has the best performance ranking with respect to an appropriate validation metric (Entekhabi, Reichle, Koster, & Crow, 2010)?

One measurement system is usually assumed a priori to be the error-free "truth" system, with other systems judged in comparison. However, the presence of inevitable errors in the "truth" system, along with differences in support scale between systems, often make the performance rankings dependent on the choice of the "truth" system, an unsatisfactory outcome. Triple collocation (TC) is a technique for estimating the root-mean-squared-errors (Stoffelen, 1998) and correlation coefficients (McColl, Vogelzang, et al., 2014) of three measurement systems with respect to the unknown true value of a continuous target variable, without unrealistically treating any one system as error-free. It has been used for estimating errors in measurements of a wide range of continuous geophysical target variables, including sea surface temperature (e.g., O'Carroll, Eyre, & Saunders, 2008), wind speed and stress (e.g., Vogelzang, Stoffelen, Verhoef, & Figa-Saldaña, 2011), wave height (e.g., Janssen, Abdalla, Hersbach, & Bidlot, 2007), precipitation (Alemohammad, McColl, Konings, Entekhabi, & Stoffelen, 2015; Roebeling, Wolters, Meirink, & Leijnse, 2012), fraction of absorbed photosynthetically active radiation (D'Odorico et al., 2014), leaf area index (Fang, Wei, Jiang, & Scipal, 2012) and soil moisture (e.g., Draper et al., 2013; Miralles, Crow, & Cosh, 2010).

Applying triple collocation to categorical target variables, however, poses unique challenges. Problems arise because categorical variables are usually unordered and bounded. As we show in Section 2, these differences mean that key assumptions in TC are violated, biasing TC

* Corresponding author.
E-mail address: kmccoll@mit.edu (K.A. McColl).

http://dx.doi.org/10.1016/j.rse.2016.01.010
0034-4257/

error estimates. In Section 3, we describe a new approach – extending the work of Parisi, Strino, Nadler, and Kluger (2014) – called Categorical Triple Collocation (CTC) that relaxes the violated assumptions and provides performance rankings for measurements of categorical variables. In Sections 4 and 5, we demonstrate its utility by applying it to the problem of ranking the performances of model, in-situ and satellite estimates of landscape freeze/thaw (FT) state.

2. Deficiencies of classical TC

Triple collocation is a commonly used technique for estimating the mean-squared error MSE (Stoffelen, 1998) and correlation coefficient r (McColl, Vogelzang, et al., 2014) of a measurement or model estimate with respect to the unknown true value of the target variable. It requires observations of the target variable from three collocated measurement systems that are linearly related to the target variable. The error model is given by

$$X_i = \alpha_i + \beta_i T + \varepsilon_i \tag{1}$$

where X_i (for $i = 1,2,3$) are the observed measurements from the noisy measurement systems, T is the unknown true value of the target variable, ε_i is a zero-mean random error term and α_i, β_i are calibration parameters. X_i, ε_i and T are all random variables. It is further assumed that $\mathrm{Var}(\varepsilon_i)$ and $\mathrm{Var}(T)$ are fixed and do not vary in time. The same assumption is not strictly required for $\mathrm{E}(T)$, although many TC studies use climatological anomalies so that $\mathrm{E}(T)$ is approximately stationary. The three measurement systems used in the analysis could be, for example, a satellite retrieval, a model estimate and an in-situ observation of the target variable. To apply triple collocation, two additional assumptions must be satisfied:

(R1) the random errors between different measurement systems must be uncorrelated with each other (i.e., $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$).
(R2) the random errors must not be state-dependent and must be uncorrelated with the target variable (i.e., $\mathrm{Cov}(\varepsilon_i, T) = 0$).

Classical TC suffers from several deficiencies when applied to categorical variables, arising from the facts that they may be unordered and/or strongly bounded. First, the additive, zero-mean error model implicitly imposes an ordering and is inappropriate for nominal (i.e., unordered) categorical variables. Second, even if we only consider ordinal (i.e., ordered) variables, the distribution of ε_i must depend on T to ensure that X_i does not take on values outside the bounded domain. This dependence violates (R2) and becomes more significant as the number of possible values the categorical variable may take on (i.e., the size of its support) decreases. Consider the case of binary variables, which only have two elements in their support (i.e., $X_i, T \in \{-1,1\}$). As shown in Appendix A, this limited support induces non-negligible correlations between the errors and target variable such that (R1) and (R2) are always strongly violated for the binary case. In particular, defining P_i to be the probability of an error occurring in measurement system i, we have

$$\mathrm{Cov}(\varepsilon_i, T) = -2P_i \tag{2}$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 4P_iP_j\mathrm{Var}(T) \tag{3}$$

which are non-zero for all non-trivial cases where $P_i > 0, P_j > 0$ and $\mathrm{Var}(T) > 0$. The observation that $\mathrm{Cov}(\varepsilon_i, T) < 0$ for categorical data has been widely noted in the econometrics literature in terms of 'mean reversion': errors tend to be biased towards the mean (e.g., Kapteyn & Ypma, 2007). The correlation between the errors and the truth then induces correlations between errors in the different measurement systems. These violations of (R1) and (R2) result in biased triple collocation error estimates.

3. Triple collocation for categorical variables

The flaws in classical TC when applied to categorical variables motivate the development of a new approach that uses an error model appropriate for unordered variables, and allows the errors and truth to be correlated. In this section, we will introduce a variant of TC for categorical variables that estimates performance rankings of three measurement systems with respect to a binary validation metric, the "balanced accuracy"

$$\pi = \frac{1}{2}(\psi + \eta) \tag{4}$$

where ψ is the measurement system sensitivity (i.e., the probability of the measurement being correct when the truth $T = 1$) and η is the measurement system specificity (i.e., the probability of the measurement being correct when $T = -1$). Unlike the simple accuracy metric μ (i.e., the probability of the measurement being correct over all cases), π avoids overestimating the quality of performance of biased classifiers on imbalanced datasets ($\mathrm{E}(T) \neq 0$), while still reducing to μ for balanced datasets. For example, consider a binary classifier which is biased, in that it always returns a classification of 1. If T is almost always 1, the biased classifier may still receive a high simple accuracy, even though it has no real predictive skill. In contrast, the balanced accuracy will more heavily penalize the classifier for the rare occurrences where $T = -1$ and the classification is incorrect. It is impossible to derive the actual balanced accuracy for each measurement system but, as will be shown, our approach allows calculation of a quantity that is proportional to the balanced accuracy for each measurement system. The relative sizes of this quantity between the three measurement systems can be used to determine relative performance rankings.

To handle unordered variables, instead of the linear regression framework adopted in classical TC, we use a classification framework. For each measurement system i and category k, define a binary classifier

$$X_i^k\left(T^k\right) = T^k + \varepsilon_i^k \tag{5}$$

where

$$T^k = \begin{cases} 1, & \text{if the true value belongs to class } k \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

and

$$X_i^k = \begin{cases} 1, & \text{if the measured value belongs to class } k \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

with $\varepsilon_i^k \in \{-2,0,2\}$, and dependent on the value of T^k to ensure that X_i^k does not take on a value outside the set $\{-1,1\}$. We may then assess the performance of the measurement system separately for each category. For instance, say we are validating a landcover type categorical variable, with the categories 'grassland', 'forest', 'desert' and 'other'. We can treat this as four different binary classification problems: 'grassland' vs 'not grassland', 'forest' vs 'not forest', 'desert' vs 'not desert' and 'other' vs 'not other'. This will result in four separate rankings for the four different categories. As a consequence, for example, the measurement system that is ranked the highest for 'grassland' may be ranked the lowest for 'desert'. There is no single, obvious way to combine these different rankings into a single ranking across categories. This is a general problem common to all categorical classification techniques. Hence, the problem of validating general categorical variables reduces to that of validating binary variables; we now drop the k superscript in our notation for convenience.