Handling Uncertainties in Image Mining for Remote Sensing Studies

Alfred Stein
ITC International Institute for Geoinformation Science and Earth Observation, PO Box 6, 7500 AA Enschede The Netherlands. Email: stein@itc.nl

Abstract. This paper presents an overview of uncertainty handling in remote sensing studies. It takes an image mining perspective and identifies ways different uncertainties. It starts with the pixel and through object identification and modeling proceeds towards monitoring and decision making. Methods presented originate both from probability and fuzzy logic based approaches.

Keywords: image mining, remote sensing, statistics

1. Introduction

Remote sensing images are used throughout as input in geographical information systems, and for spatial decision making. These images are increasingly available, with an increasing resolution in space and an increasing temporal resolution. But as the questions posed towards these images are changing as well, A request remains to quantify the uncertainties.

The common way of using remote sensing images is to start with the process that we wish to observe. Such processes (like the state of the earth surface at some time point) can sometimes be observed with a single image. Typical examples include environmental pollution, biodiversity assessment and inventorization of urban quarters. Other processes require a series of images to be observable. These we encounter in monitoring flooding, forest fires and urban developments. On the basis of pixel information (or radar and/or lidar signals) objects are identified and subsequently classified. In monitoring studies, the objects are subsequently tracked in time, in which their behavior may be governed by external factors that have to be identified and quantified. The future (and the origin) of processes most likely have to be predicted. The quality of both the modeling and the predictions is influenced by the quality of the images, i.e. the resolution in space and time, the sacral characteristics of the signals, but also by the crispness (or fuzziness) of the objects. For a proper analysis of the quality of the objects, therefore, an analysis of error propagation is important. Finally, the modeled objects and their predictions can be used for decision making. We distinguish here both decisions made on the basis of single instances, like a farmer who has to an immediate action when observing e.g. crop stress, as long-term decision making on a series of images, like in global change issues.

Error propagation within a GIS context has been done in the past [1], [2], leading to assessment of uncertainty of the derived objects, as a polygon or as a raster. Studies in the field of remote sensing images have been carried out as well [3], [4]. The aim of this paper is to give an overview of uncertainties existing in information extracted from remote sensing images and to sketch several ways to handling them. The aim is to focus on those aspects that currently still require attention.

2. Image mining

Image mining is defined as "The analysis of (often large sets of) observational images to find (un)suspected relationships and to summarize the data in novel ways that are both understandable and useful to stakeholders" (Fig. 1). Objects discernable on those images can be either crisp or fuzzy and vague. The notion of crispness versus vagueness depends largely on the scale of observation. Crisp objects correspond to
well defined entities with sharp boundaries. At a 1m resolution, a building may serve as an example of a crisp object. Vague objects are either objects with a poor definition, or objects with a vague boundary. Two examples with an uncertain boundary and an often rather subjective definition are an area with crop stress in a farmer’s field and a patch with high biodiversity in a forest. Landscape examples are a dune (or a mountain) with no apparent support, a city having gradual transition zones to the rural land and a contaminated river. These types of objects all have to be extracted for various reasons from remotely sensed images.

We distinguish five important steps in image mining: identification, modeling, tracking, prediction and communication with stakeholders [5]. All these processes will be briefly discussed below. On top of this, we notice aspects of spatial data quality in each of these steps.

Identification of uncertain objects commonly proceeds by making the step from raster to objects. Usually, a segmentation of the image is done first, providing more or less homogeneous segments, followed by a classification, although pixel based classification may start with a classification, followed by grouping of the pixels that are classified into the same class. In the first approach, identification is done by applying a segmentation routine, in which both the object and the uncertainty are modeled. Various procedures for image segmentation are well documented, and include procedures based on mathematical morphology, on edge detection and on identifying homogeneity in one band or in a set of bands. Classification routines include statistical routines such as k-nearest neighbor classifiers, and increasingly also fuzzy classification methods. Modeling of uncertainty has been done in the past by using, e.g. a confusion index, whereas traditionally a discriminant analysis was applied, honoring the different spectral bands and yielding posterior probabilities to the non-selected classes. Currently, interesting results are being obtained by applying a texture based segmentation, based upon single bands and or multiple bands. This leads to improved segmentation, again including uncertainty values. The result of this operation for an image at moment $t$ thus is a series of $n_i$ objects $O_{t,i}$, $i = 1, \ldots, n$, that are characterized by similar pixel values, which are different from pixels values in the vicinity.

Modeling of identified objects $O_{t,i}$ can typically be done applying a fuzzy approach, resulting in membershipfunctions of the objects to a class $A$ of interest. Membershipfunctions in a 2-dimensional space are functions $\mu_A(x,y)$, taking values between 0 and 1, which specify the degree to which the location $(x,y)$ is characterized by $A$. Such membershipfunctions are characterized by the steepness of their slopes, i.e. showing how rapid it increases from zero to one, by their homogeneity and by their support and centroid.

![Image mining for vague objects.](image.png)
3. Source of error 1: the pixel

3.1. Resolution and location

We consider a digital image as an array \( A \) of numbers. Each of these numbers \( \{a_{ij}\} \) has a row index \( i \) and a column index \( j \). In principle, these numbers coincide with a spatial location \( s \). One of the main issues to be dealt with first concerns the use of a set of ground control points that should suffice to tie the image within reasonable limits to the earth surface. As optical and radar remote sensing, however, are aggregated reflection values, the place location is leading to uncertainty. Moreover, uncertainty increases as the earth surface is curved (which may not be negligible in large area studies) and the earth surface contains differences in elevation. This third dimension may have a profound effect on the quality of the coordinates. As a start, one may assume that the coordinates contain an uncertainty \( dx \) in the \( x \) direction and \( dy \) in the \( y \) direction, e.g. corresponding with the image resolution. These uncertainties have an effect on the points of gravity of the objects, having an uncertainty equal to the standard error, i.e. \( \frac{dx}{\sqrt{n}} \), with \( n \) being the number of pixels included in the object.

3.2. Digital number

Next we consider the digital number as such. In each of the visible bands we have numbers ranging from 0 to 255, that means of combination can be changed into an index that we know is well related to specific aspects of the land cover. Typical examples are the NDVI for vegetation and the NDSI for snow cover. The digital numbers, however, are affected by various distortions between the earthy surface and the sensor. Typically, an average value is taken, leading to a representation of the cover of an area approximately equal to the spatial resolution of the sensor. This may contain a uniform area, but may also reflect various forms of land cover within the single pixel resolution. Between the earth surface and the sensor atmospheric distortion may be present, that may lead to noise in the signal. Thereafter, physical characteristics of the sensor may lead to further noise, hence causing the registered value to be different from the condition of the earth surface. When images are used for monitoring, this may be even more of a problem as differences in land and air conditions may differ within single days as well as between different days of registration.

So far, we have focused on both positional accuracy and on attribute quality. An important phenomenon in radar earth observation, however, is the presence of speckle. Speckle originates from interference of coherent responses coming from many scattering elements within a resolution cell. It results into a large variance of radar image compared to its mean value, therefore SAR images appear noisy. Speckle differs from random noise, as repeating a radar image acquisition in the same conditions reveals exactly the same speckle pattern. In an image analysis the presence of speckle is a complicating factor, to some degree similar to noise for optical images. As a result of speckle, phase values in radar images are not informative in most cases. An important exception is radar interferometry, which makes use of the difference between the phase values in two images.

4. Source of error 2: the objects

The objective of many remote sensing studies is to identify objects that have an ontological representation on the earth surface, We denote these with \( \{O_i\}_{i \in I} \). These objects can have different meanings, and they can be of various different types and shapes. Commonly, a segmentation procedure is applied first, identifying homogeneous (in some sense) sets of pixel values, in one or more bands. The objects can have sharp boundaries, but these boundaries are not necessarily clearly identifiable by satellite sensors, for example due to the limited resolution of the bands, and by the noise produced as identified earlier. Other objects may be less precise and less clearly defined, causing gradual and/or indefinite boundaries. This leads to the well-known issues of spatial data quality, basically referring to objects in geographical information systems.

The modeling of objects is the next step to take. The aim of the modeling of objects is usually to be able to relate them to other sources of information, like for example a topographic map or a digital elevation model. One may wish, for example, to follow the objects in time, or one may wish to properly classify them, resulting into landuse (in stead of land cover) units. The ways that are commonly used for modeling are the crisp objects, for which polygons are suitable, and fuzzy or vague objects that can be modeled with membership functions from a fuzzy logic perspective. In the modeling stage, several sources of error may
emerge. When a polygon based modeling is done, the boundaries may ignore the uncertainty on the objects, an apparent angular boundary may be simplified to a smooth boundary. On the other hand, a fuzzy object may lead to cumbersome operations (like determining its size and shape or combining the objects with existing topography) and hence may prohibit an efficient handling and thus lead to reduced functionality when it comes to further stages. One of the problems to deal with concerns the storage of uncertain objects: the fuzzy based method usually requires a raster based approach, whereas in particular topographical information is largely vector based. Changing a fuzzy object to a vector representation requires a large number of vectors, and hence long computation times. This large number of vectors in turn leads to pseudo exactness, which in fact entails another approximate reasoning: the number of vectors should in principle be infinite (one for each membership value), but in actual GIS calculations this number is reduced to a convenient number.

An issue already touched upon above concerns the issue of small objects. Small objects, or linear features may have an effect on the digital number registered by a sensor but may not be identifiable in geometric terms. Recent developments in super-pixel resolution have shown that such objects can be identified with some reliability. The basic idea is that a particular geometrical shape is present, which may then lead to the identification. An oil pipeline, for example, may be visible in some parts of an image, whereas one may assume a linear connection between those objects. This leads on some bands to deviating pixel values where the road is present, whereas the other pixels do not have such effects. One way of proceeding thus is by using linear (and unlinear) unmixing procedures. Superpixel resolution mapping, however, goes somewhat further, as it allows us to identify a range of different objects, not a prior given.

An interesting step concerns the validation of objects. Crisp and uncertain objects require a validation by means of ground truth. When objects are uncertain, such validation is a somewhat complicated activity. Even on the ground, the delineation of such objects (like a city) may be impossible, as it does not have clear and unambiguous boundaries. Also the presence of this city on a MODIS 250m resolution image may be impossible, as transition boundaries are likely to be present. Hence finding the correspondence between these objects requires an overlaying of uncertain concepts. The only way apparently would be make an abstraction of the concept ‘city’ as a representation within an information system, thus allowing for uncertain and gradual boundaries. On the one hand, such an abstraction is likely to be sensitive to subjective influences, whereas on the other hand also the overlaying of two such objects (one from the MODIS image and one from the representation) is not a trivial task, although it can be carried out.

In other field of science it brings us to the important issue of sampling: the question has been often raised and as far as I know only partly been answered, how a dense the optimal sampling configuration should be for validation. A commonly applied tool in remote sensing classification is the error matrix, containing both the user and producer accuracies. A limited number of samples justifies the classification, as expressed for example by the kappa statistic. Clearly, each identified class should at least have one field sample for validation, whereas it would be better to have a set of samples. But identifying vague objects in the field is not a straightforward matter. The optimal number, nor the optimal allocation of these is at present fully and in all generality worked out. More emphasis has been given to the sampling (and subsequent interpolation) of identifiable objects in environmental studies. Recent work has tried to identify objects of oil contaminated soil in Nigeria. A spatial interpolation method (Bayesian Maximum Entropy) has been applied to deal with the minimum number of samples. It showed that it better performed than any form of kriging, and an empirical relation with soil EC proved to be of great value. Clearly, the final word should yet to be spoken on this matter.

5. Source of error 3: the monitoring

The next step in complexity concerns the monitoring of processes. Remote sensing has been applied successfully to monitor changes in glacier coverage, forest fires, changes in urban sprawl and changes in vegetation, to mention a few. These processes can be monitored if a number (2 or more) images shows the same phenomenon at different time steps. First, difficulties in storing and representing these objects have to be overcome, as they are governed in principle by all the of the uncertainties described above: the objects may be inherently vague and uncertain, their registration may be uncertain, they may only be partly visible,
problems in coordinates may exist, etc. But it may be the case, like in examples from flooding, that the problems can be solved to a manageable small number, leaving us with enough confidence to proceed. Typical questions to answer would then be the speed of changes, the occurrence of sudden events and the origin and future of the objects. The analysis carried out so far will thus get a temporal dimension. Clearly, the temporal scale of observation should match the scale of changes in the earth process. It is of little use to observe the continental drift with METEOSAT images, whereas MODIS data are probably of little value when we wish to analyze the movements of a forest fire. In principle, the monitoring faces similar problems in terms of its uncertainty as the single images, with the additional complexity that we need a third coordinate, namely time.

6. Source of error 4: predictions

If we have sufficient observations to monitor the process, we may wish to make predictions of the process towards the future. That would mean that we predict some parameters (like the point of gravity, or a shape parameter) towards any point in the future. Statistical time series analysis methods may be of use here. An alternative way is to make simulations of the process towards the future. This usually requires a set of assumptions, but it allows us then to compare different scenarios. Making predictions requires us to expand beyond the domain of observations. This is always tricky and may lead to additional errors. Statistical procedures from time series analysis or even a simple regression model are indispensable to quantify these uncertainties.

7. Error propagation

7.1. The error model

To describe a statistical error model, we consider a range of data \( z_i \) of \( n \) repeated observations, and consider the mean, \( m_Z \) as a summary measure. This is valid if the data are a realization of random sample of a single stochastic variable, \( Z \). An exact distribution of \( Z \) is most often not required, because of the central limit theorem. Also, it may not be really necessary, as statistical interpretation tools are robust. However, a careful testing should be done to see whether the data are from a random sample. Several issues can be addressed by using box plots, such as the presence of outliers, or the occurrence of systematic change.

To set up a full error mode, we suppose that there is a true value \( \zeta \) and that \( X \) is a measurement variable for this value. This variable is supposed to have a finite expectation, the so-called measurement expectation: \( E[X] = \mu \). The measurement variance is then given by \( E[(X - \mu)^2] = \sigma^2 \). We distinguish between the systematic error (the bias), being the difference between \( \mu \) and \( \zeta \); \( \delta = \mu - \zeta \), and the random error: \( e = X - \mu \). Combining this we see that the measurement variable is the sum of true variable, the bias and the random variable:

\[
X = \zeta + \delta + e \tag{1}
\]

We notice that the common assumptions apply, namely that \( E[e] = 0 \) and \( \text{var}[e] = \sigma^2 \). It depends on the circumstances whether \( \delta \) should be taken into account; if \( \delta < \frac{1}{2} \frac{\sigma}{\sqrt{n}} \), then the random errors dominate the mean, and \( \delta \) can be ignored. Also if we consider differences only between two variables with the same \( \delta \) then \( \delta \) can be ignored as it disappears from the differencing. In many studies, systematic and random errors can not be totally separated.

Measuring the quality is done by assessing the discrepancy between two datasets in various ways, depending upon available world data. Typically, use is made of an independent test set, an independently selected sub-set, or a leave-one-out procedure. If it is done in this respect, then we compare observed \( (x_i) \) data with real world \( (t_i) \) data, e.g. using one of the following measures:

- ME: Mean error: \( (1/n) \sum (x_i - t_i) \) \tag{2a}
- MSE: Mean squared error: \( (1/n) \sum (x_i - t_i)^2 \) \tag{2b}
- MAE: Mean absolute error: \( (1/n) \sum |x_i - t_i| \) \tag{2c}

If a large bias exists, then the ME is large with respect to the actual measurements, whereas if a large uncertainty exists then both the MSE and MAE are large.

7.2. Propagation of error

Next, one may question how important the uncertainty in the data is. Small errors may be negligible in
subsequent calculations, whereas large errors may have a large impact. This question is then to be answered by means of an error propagation [1]. To do so, we assume that a given function $f$ exists, that in this modeling is considered to be a continuously differentiable function of $n$ observable variables: $\zeta_1, \zeta_2, ..., \zeta_n$. The function is then given by $z = f(\zeta_1, \zeta_2, ..., \zeta_n)$. Also, there are $n$ observables $X_i = \zeta_i + \delta_i + e_i$ with $E(e_i) = 0$ and $\var(e_i) = \sigma^2$. As the aim is to estimate $z$, we need an estimator $Z^*$, as well as the variance of $Z^*$. As a first guess for $Z^*$ we take $Z^* = f(X_1, X_2, ..., X_n) – \zeta$ we therefore measure $X_1$, ..., $X_n$ once, and determine $Z$ with $f$. But as we may not have taken the measurements in the point $(\zeta_1, \zeta_2, ..., \zeta_n) = \zeta$, we determine a Taylor series around $\zeta$. This expansion is given by

$$Z = f(\zeta) + (X_1 - \zeta_1)f_1 + (X_2 - \zeta_2)f_2 + ...$$

$$\frac{1}{2} (X_1 - \zeta_1)^2 f_{11} + (X_2 - \zeta_2)(X_3 - \zeta_3)f_{12} + \frac{1}{2} (X_2 - \zeta_2)^2 f_{22} + ...$$

where $f_i$ denotes the derivative of $f$ according to $\zeta_i$, and $f_{ij}$ the 2nd order derivative of $f$ according to $\zeta_i$ and $\zeta_j$. Both first and second order terms should exist. Equation (3a) can also be written as:

$$Z = f(\zeta) + (\delta_1 + e_1)f_1 + (\delta_2 + e_2)f_2 + ...$$

$$+ \frac{1}{2} (\delta_1 + e_1)^2 f_{11} + (\delta_2 + e_2)(\delta_1 + e_1)f_{12} + \frac{1}{2} (\delta_2 + e_2)^2 f_{22} + ...$$

The second expression is in particular appealing as it contains the systematic and the random errors. In various cases, we may be able to make various assumptions. The most common are:

- No bias occurs, i.e. $\delta_i = 0$, then $z = f(\mu)$
- The $X_i$ are uncorrelated, hence $E(X_iX_j) = 0$
- The probability of large accidental errors is small, hence 2nd and higher order terms can be ignored in the Taylor series

This then brings us to the two famous laws of error propagation:

The law of the propagation of random errors, valid if all three assumptions hold, equals

$$\var(Z) = f_{11}\sigma_1^2 + f_{22}\sigma_2^2$$

The law of the propagation of bias equals

$$\mu_Z \approx f(\mu) + \frac{1}{2}(f_{11}\sigma_1^2 + f_{22}\sigma_2^2)$$

This law shows that systematic errors (bias) may occur as a consequence of random errors: $\delta Z \approx \frac{1}{2}(f_{11}\sigma_1^2 + f_{22}\sigma_2^2)$.

### 7.3. A simple example

A simple example is the area of an ellipse. It is well known that area of an ellipse is equal to $\text{Area} = \frac{1}{4}\pi \zeta_1 \zeta_2$. Here $\zeta_1$ is the length of one axis, and $\zeta_2$ is the length of the other axis. Notice that the area of a circle is a special case, with $\zeta_1 = \zeta_2$. We may then question what the error propagation would likely be, given uncertainty in the two coordinates, $\zeta_1$ and $\zeta_2$. In order to make quantitative statements, we consider the function by $z = f(\zeta_1, \zeta_2) = \frac{1}{4}\pi \zeta_1 \zeta_2$. We first calculate the two derivatives, that are required in equation (3a), and find that $f_1(\zeta_1, \zeta_2) = \frac{1}{4}\pi \zeta_1$ and that $f_2(\zeta_1, \zeta_2) = \frac{1}{4}\pi \zeta_2$. Both derivatives are therefore linear functions in one of the coordinates. Proceeding along these lines, we find that two of the second derivatives are equal to 0, namely the derivatives according to the same variable twice, whereas $f_{12}(\zeta_1, \zeta_2) = \frac{1}{4}\pi$, a constant. This is also the value that emerges for $f_{21}(\zeta_1, \zeta_2)$.

![Fig. 2. The area of an ellipse, as a function of the length two axes $\zeta_1$ and $\zeta_2$.](image-url)
Now suppose that we are interested to find the value in a few points. First we take the point \( z = (1,2) \), and suppose that we observe the values \( x_1 = 1.1 \) and \( x_2 = 2.1 \). For these values, \( f(\zeta_1, \zeta_2) = \frac{1}{2}\pi \). We then find for the area the value of

\[
z = \frac{\pi}{2} - 0.1\cdot\frac{1}{4}\pi\cdot 2 - 0.1\cdot\frac{1}{4}\pi\cdot 1 + 0.1\cdot 0.1\cdot\frac{1}{4}\pi
\]

\[
= \frac{\pi}{2} - 0.075\pi + 0.0025\pi
\]

A similar calculation can be done for other values of \( \zeta_1 \) and \( \zeta_2 \), possibly leading to different values of the area and errors in its calculation.

8. Communication

Communication of uncertainties turns out to be a major challenge in remote sensing studies. Apparently, the user has to make a single decision in the end, and although uncertainty based decision making is a well-known scientific subject [8], [9], its communication has somehow never reached wide-spread acceptance. Recent discussions have revealed that at best statements like ‘An area has a low probability of a specific property’ or ‘An area has a high probability of that property’ can be communicated to the general public. A proper appreciation of critical values like 0.05 and 0.95 is still almost impossible. Clearly, this topic has to be further elaborated upon to receive general acceptance.

9. References


