Yield Monitors and Remote Sensing Data: Sample Statistics or Population?

Terry W. Griffin, Raymond J.G.M. Florax, and Jess Lowenberg-DeBoer

At precision agriculture data analysis workshops there are often diverging views on the kind of statistics that is appropriate for sensor based or remotely sensed data such as yield monitor data and imagery. One view is that yield monitor data represent a sample. In that case, some sort of sampling theory applies and well-known tools from spatial statistics and geostatistics should be utilized. The concurrent view is that precision agriculture data on yields actually represent the population; the use of sample statistics is therefore not pertinent and simple averages or proportions referring to this one state-of-the-world can be calculated without having to evoke statistical inference. At first sight, it may seem natural to believe that taking simple averages from treatment blocks and comparing these to other averages provides useful decision making information. Many advocates of the idea that precision agriculture data are population data, farmers and researchers alike, argue that simple averages of yield monitor data are sufficient for whole-farm decision making purposes. Unfortunately this is not the case. Standard farm-level software therefore often includes tools that summarize yield or other site-specific data into classes according to soils or other predefined management zones.

Although there are many degrees of freedom with site-specific yield monitor data, this does not imply that there is no sampling, and hence no stochastic process involved. With thousands of observations, and thousands of repeated measurements for each treatment, there is likely to be a negligible difference between statistical inferences based on using sample counts as compared to the "true" population value. The large number of degrees of freedom makes sample estimates very precise. For instance, if there were 20,000 observations, the differences in mean squared error or many other statistics is roughly in the order of 1/20,000. This does, however, not imply that the population view is necessarily correct, and that simple averages across treatment blocks can be determined and compared in a deterministic fashion. Two reasons for why simple averages of precision agriculture data are not sufficient for whole-farm management design are discussed below.

Data collection resolution matters

Although very dense yield data or even a remotely sensed raster cover nearly the whole field, data has been aggregated into yield points or imagery pixels. Harvesters are most frequently set to collect yield measurements on one-, two-, or three-second intervals and imagery of less than a meter are becoming commonplace. Regardless whether the imagery has a resolution of 1 decimeter or 3 meter or the yield monitor records data every 5 meters, the data are not an example of infinitesimal observations eventually making up the whole population. Even if the data were infinitesimally small and we would be able to collect them all, we would still not capture the population, because changing harvester interval times would change the recorded yield measurements. The recorded attribute values are therefore just one realization of what could have been observed. The definition of the spatial object for which yields are collected can be varied, and some type of sampling theory is therefore appropriate.

The modifiable areal unit problem (MAUP)

The starting position and direction of the same harvester may provide different yield measurements. For instance, if instead of harvesting the first six rows and then harvesting the subsequent six rows throughout the field, six rows were harvested beginning three or four rows from the edge of field, different yield measurements would certainly be made even though there were six full rows of crop harvested in the same general area. This is an example of the modifiable areal unit problem (MAUP) discussed in the spatial sciences and statistics. MAUP has to do with the size, shape, and orientation of a grid superimposed over an area. If the harvester passes were instead perpendicular or at a diagonal to the abovementioned direction, the yield measurements would be different due to a number of factors including different plants being harvested for a given yield measurement and associated harvester dynamics, potentially impacting statistical results. Holt et al. (1996, p. 181) state "if a statistic is calculated for two different sets of areal units which cover the same population, or sample, a difference will usually be observed even though the same basic data have been used in both analyses. This difference is the modifiable areal units problem."

A given field is a sample of possible fields

It may be tempting to now argue that there are of course other spatial configurations possible for which we could have collected the data, but still total yield is total yield, and in that sense we do observe the population, and not merely a sample. In other words, the spatial objects may vary, but the attribute value of the spatial objects (in this case, yields) is an observation of the one-and-only existing state of reality. However, the data in question are only one realization of how that field could have been farmed in that year. So, there is an intriguing notion that statisticians call "superpopulation" which refers to other possible states of the world that *could* have been observed (Haining, 2003, pp. 51-54). We could have collected the yield monitor data one week earlier, or one week later, and they would have been slightly different. Or, the actual weather could have been more or less conducive to crop growth and ultimately have resulted in different yields. Hence, a given field that we observe is but a sample of many possible fields that could have been observed.

What sampling theory?

From the above we can deduct two reasons why the population view of precision agriculture is at odds with reality and with what is desirable. In precision agriculture research we are not dealing with a deterministic world for which we can make perfect predictions. For one, the way in which we collect data is not deterministic. Implicitly, we define *spatial objects* that could have been defined differently. But even if that were not true, we observe yields for just one field. The *attribute values*, yields in this case, could have been different if we had observed them at a different point in time or under slightly different circumstances (less sunlight, more rain, etc.).

Conclusions

If yield monitor and remotely sensed attribute values were deterministic, meaning that each random draw of the distribution yielded identical results, then simple treatment averages of exhaustive data, i.e. extremely high resolution, may be sufficient for decision making.

However, biological systems in general and agricultural production attributes in particular are stochastic, meaning that any given draw from the distribution may result in differing outcomes even in a given year. This means that sample statistics must be used to make reliable decisions from yield monitor and remotely sensed data. The type of sampling theory that is most appropriate should be determined on the basis of whether we view the spatial objects and the attribute values as deterministic or stochastic. For instance, in an analysis of diseased crop both the location (spatial object) and the attribute values (disease or no disease) are stochastic. In other cases, one of the two may be stochastic and the other deterministic. One can have an extensive and interesting discussion about the choice of an appropriate theoretical sampling framework. However, the population view of precision agriculture seems like a case of mistaken identity.

References

Haining, R. 2003. *Spatial Data Analysis: Theory and Practice*. Cambridge, UK: Cambridge University Press.

Holt, D., Steel D.G., and Tranmer, M. 1996. Area Homogeneity and the Modifiable Areal Unit Problem. *Geographical Systems*, 3, 181–200.