

Mapping farm survey data in rural and regional Australia

Caroline Rasheed and Teresa Neeman
Australian Bureau of Agricultural and Resource Economics

Mapping Science Conference
Sydney, 3–6 December 2000

ABARE project 1632

ABARE

Innovation in Economic Research

GPO Box 1563 Canberra 2601 Australia
Telephone +61 2 6272 2000 • Facsimile +61 2 6272 2001
www.abareconomics.com

Introduction

ABARE has been conducting annual farm surveys for more than fifty years. The survey samples approximately 2 per cent of broadacre and dairy farms with an estimated value of agricultural operations (EVAO) of at least \$22 500. The 2 per cent sample represents approximately 70 000 broadacre and 13 000 dairy farms across Australia. The data collected provide a profile of the physical characteristics of the farms, farm management practices, farm financial performance and a profile of the farm family. Supplementary surveys conducted at regular intervals collect additional information on farm management practices and farm lifestyles, including spending patterns, technology adoption and attitudes toward change.

The annual farm surveys are designed to provide estimates of average farm performance at state and industry levels. State and industry level averages of farm physical characteristics and financial performance are published in tabular form each year. However, maps of these data may have a greater visual impact than data presented in tables. Maps can show variation of averages within regions, as well as between regions. Differences between and within regions are easily visualised through the prudent use of color contours, and may provide a framework for further research into the reasons behind these differences. Maps generated serially across many years of the survey may show the evolution of rural industries or the impact of climatic conditions on farm performance. Maps also provide a visual framework for exploring associations between land management practices and farm performance.

Maps are not just visual images of average values of farm performance. A map showing measures of variation may reflect the diversity of farm performance and farming activities in local areas. Diversity can be the result of large variations in climatic conditions over time, which result in variable crop yields. Diversity over time can also be the result of changing demands for commodities and changing prices. Mapping these measures of diversity may also provide a comparison of the adaptability of farming industries between regions.

There are important considerations when creating a map based on a survey of the population. Like any graphical display, choices of scale and color can dramatically influence the interpretation and the interpretability of the map. An additional complexity of mapping arises with the wide range of modeling techniques available. For ABARE's farm survey data, the initial challenge is to extend the information collected from a representative sample of the farm population across the whole of the agricultural regions of Australia, to give a picture of local and regional variation in the farm population. The extension of mapping data from individual farms to reporting local averages across agricultural regions is a process called smoothing. Information collected at a sample farm may reflect in part what is happening on nearby farms in the area. Having a clear understanding of the extent

to which a sampled farm is representative of other farms in the area is critical to producing a reliable map.

There are many ways to smooth or model data; the ultimate appearance of the map is sensitive to these choices. Figures 1 and 2 are examples of smoothing processes that estimate average farm area across Australia. In each case, a simple smoothing algorithm with a constant smoothing parameter was used. In the first figure, local averages are reported in areas where the farm density is high. Farms tend to be smallest in the high rainfall zone along the coastal belt of the eastern states of Australia. The steep topography makes this area less suitable for large scale cropping, while richer soils and adequate rainfall mean that more intensive smaller scale farming is possible. Also, because these farms tend to be in less remote areas, higher land values may constrain farm size. Farms are, on average, larger further inland in the wheat–sheep zone where the climate and topography allow large scale cropping and grazing. In the Northern Territory, where farm density is low, the smoothing process does not extend to give local estimates outside areas with low or no sample numbers. Ironically, the Northern Territory is a region that is heavily sampled; almost a quarter of the farm population responds to the annual survey.

Figure 2 illustrates what can happen when the farm data are smoothed over larger areas. All of the detail around the coastal area is lost, because the larger farms in the wheat–sheep zone have contributed to the estimates of farm area along the coast. However, because sample values are being averaged across wider areas, the sample taken in the Northern Territory is more informative for the whole of the agricultural region there.

Neither map gives a satisfactory representation of the data. In the following sections, some of the alternatives to the simple smoothing procedure that remedy some of the problems seen in these two figures are outlined. In the following section, some of the goals in smoothing and how different approaches can be used for different goals are described. In the third section, a modified smoothing algorithm used to create maps is described. Further discussion centres on issues of reliability of estimates and maintaining confidentiality of individual farm data.

The modeling process

There are many ways to model data; the method chosen depends on the purpose behind the presentation, as well as on the data themselves. In all cases, data are collected at selected points or areas, and the goal is to estimate a value at every point in Australia. For some purposes, the best estimator at a point is the measure at that point. For example, if rainfall were to be used in a model for predicting crop yields, then data collected at a station is a better estimator of rainfall in that immediate area than local rainfall averaged from data at stations 100 kilometres away. If no measure was taken in that immediate area, then a weighted average of nearby points may be a best guess. For other purposes, preserving

Figure 1: Broadacre farms, 1989-99: Area operated – all

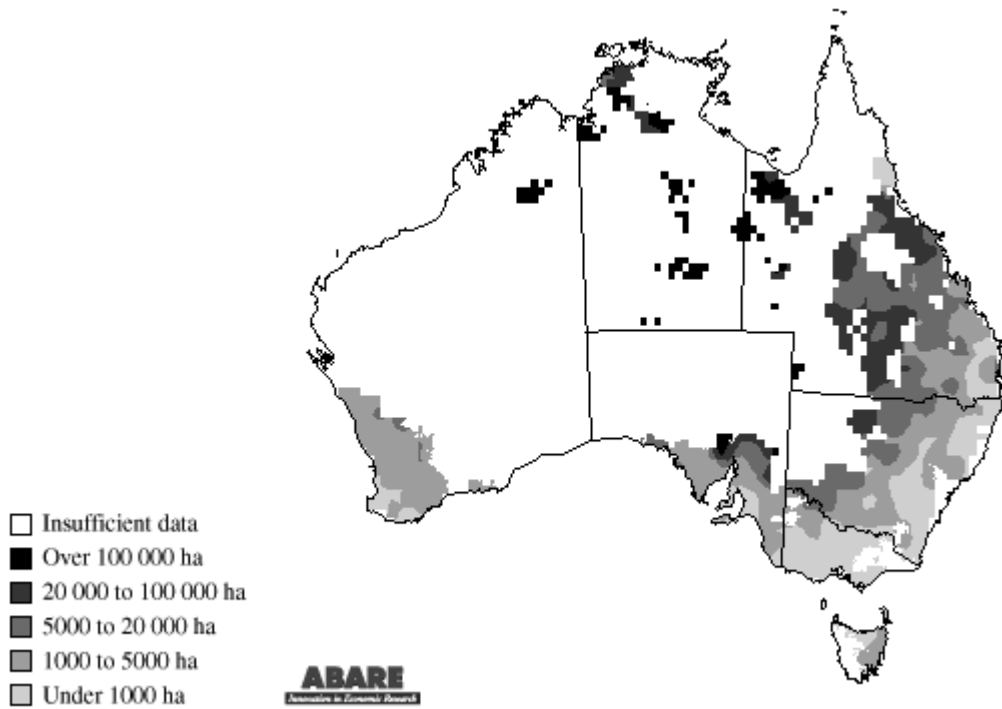
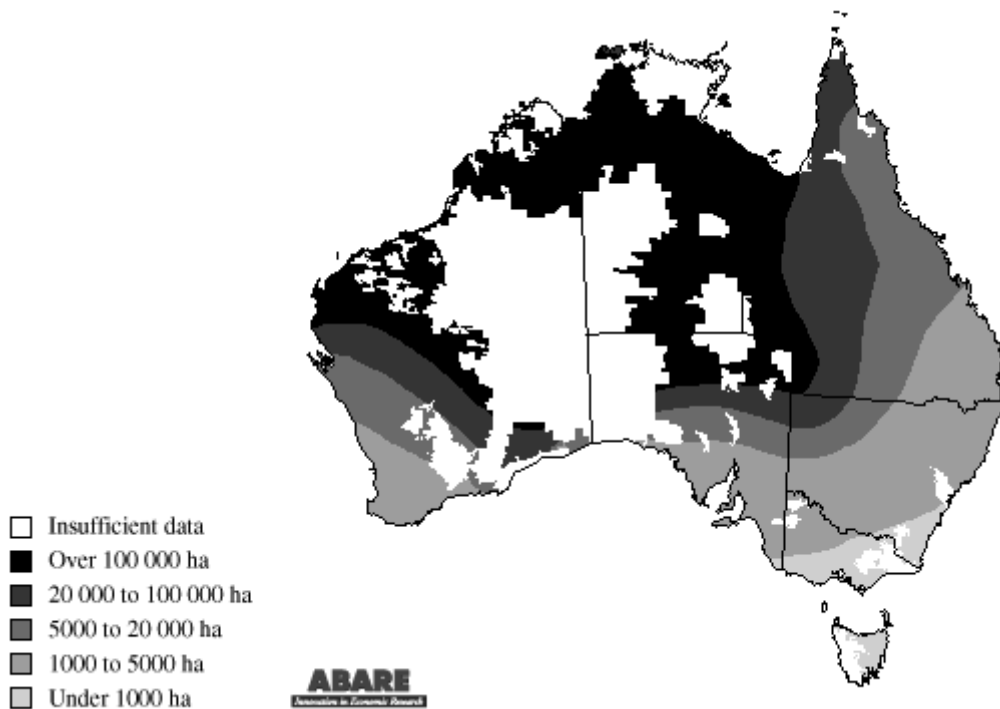


Figure 2: Broadacre farms, 1989-99: Area operated – all



the point measure may not be optimal. If rainfall maps were to be created to compare average rainfall patterns between years, or to compare average rainfall across regions, then a model that smooths out the extreme local or temporal variability may be better at identifying overall general trends and patterns.

When reviewing farm financial performance or crop performance, the emphasis is not on identifying individual farm values, but rather on identifying areas affected by unusual seasons, or adverse or favorable marketing conditions. Point prediction is unimportant in this situation, while understanding local and regional trends and variations is important. In fact, precise information collected from individual farms should be kept confidential even when area averages are reported. Unlike rainfall data, which can in theory be collected at finer and finer detail, the amassing of farm data ends with a complete enumeration of the farms. If geographic coordinates were available for each farm as well as measures of farm performance, then these measures could be mapped. Variations from farm to farm in a small area would give these maps a speckled look, and hide general trends. Taking local averages of census data could bring out these trends, and give a greater visual impact.

Mapping local averages, however, gives only part of the picture. As with any data summary, reporting the local average together with a measure of variability around that average value gives a feel for the degree of diversity in the region.

An essential part of producing reliable maps is understanding what is meant by 'local' in the term 'local average'. Local could mean averaging together farm data within 10 kilometres or within 300 kilometres. It could mean averaging together farms that are within a certain distance of a point, or only averaging farm values within a certain distance that also lie within a certain region. 'Local' can mean something different in the Northern Territory, which is sparsely populated with farms running cattle on huge tracts of land, compared with the south eastern corner of Australia, which has a very diverse mix of irrigated and dryland farms, sheep and prime lamb farms, dairies and other livestock farms. The appearance of a smoothed map is sensitive to the definition of local that is chosen.

The parameters that determine what is meant by 'local' are called the smoothing parameters. They effectively determine to what extent nearby or distant farms contribute to a local average. When the smoothing parameters are large, data from distant farms are averaged with data from nearby farms. In large areas where land use and farm characteristics are relatively homogeneous, using the more distant farms improves the reliability of the estimate of the average. However, in areas where land use patterns are very localised because of geographic heterogeneity or variable climate patterns, averaging together data from distant farms with data from nearby farms may obscure important localised patterns that are well represented by the nearby farms. In these areas, the resulting maps will be over-smoothed.

Using small smoothing parameters will result in more detailed maps. These maps give a reliable picture of farm diversity in regions where the sample density is high. However, it should be recognised that the smaller the smoothing parameter, the fewer farms contribute to the local average at each point. This can be a problem in areas where the sample is sparse. The local average at these locations could be made more robust by including information from farms further away that are representative of farms in the area. In addition, there may be areas where an average is not available because there is no sample farm in the immediate area, even though farms further away are similar in structure.

Modeling and mapping survey data

ABARE's annual farm surveys have a stratified design; the strata are based on region, size of farm operations and farm industry. After the sample data are collected, survey weights are estimated for each farm so that totals of designated physical benchmarks are consistent with known totals from regional census figures. This survey weight approach is based on Bardsley and Chambers (1984). The sampled farms together with their survey weights represent the total population of farms in the region. Regional averages (A_R) and regional totals (T_R) for a given variable of interest (for example, total cash receipts) are then estimated using the survey weights, as follows:

$$(1) \quad A_R = \frac{\sum_{i \in s \cap R} w_i y_i}{\sum_{i \in s \cap R} w_i}, \quad T_R = \sum_{i \in s \cap R} w_i y_i$$

where $s \cap R$ denotes the set of sample farms in a given region R ; w_i are the survey weights; and y_i is the variable of interest. The area estimates obtained by aggregating smoothed data should be consistent with these design based estimates. For this reason, survey weights are a critical component in the smoothing algorithm for ABARE survey data.

Local area averages for a given variable of interest (for example, total cash receipts) are estimated using a modified Nadaraya-Watson estimator across the agricultural regions of Australia. The modifications are made to account for the survey weights w . For a given variable of interest, y , an estimate of y given a location x , is given as:

$$(2) \quad \hat{y}_x = \frac{\sum_{i \in s} w_i K_h(x_i, x)}{\sum_{i \in s} w_i K_h(x_i, x)}$$

where w_i and y_i are respectively the survey weight and the measure of the variable of interest corresponding to sample farm i , and x and x_i are location coordinates. The function K_h is the Epanechnikov kernel, defined as:

$$(3) \quad K_h(x_i, x) = \begin{cases} \frac{1}{h^2} \frac{2}{\pi} \left[1 - \frac{\|x - x_i\|^2}{h^2} \right] & \text{if } \|x - x_i\| < h \\ 0 & \text{otherwise} \end{cases}$$

In simple terms, the estimated value of y at a location x is a weighted average of values at nearby sample farms. The contribution that a sample farm makes to this weighted average depends on its survey weight w_i and its distance from location x . Farms further than distance h from location x contribute nothing to the local average at that point.

From the point of view of a sample farm, the smoothing process can be thought of as a way of spreading the sample data over an area around which this farm may be representative. The maximum distance over which a farm's values have influence is called the bandwidth, and is denoted by h . The farm's area of influence is πh^2 , although this influence should be small near the area boundary. This area of influence is, in effect, the smoothing parameter. Large areas of influence around every farm produce very smooth maps as in figure 2, whereas small areas of influence produce maps as in figure 1. In order to draw from the favorable aspects of each of these maps, a different area of influence can be set for each farm. How this area is chosen may depend on underlying geography and climatic patterns in the farm vicinity, or on particular characteristics of the farm. In practice, we allow the area of influence to depend on the survey weight of the farm and the area of the farm.

Heuristically, the higher the survey weight of the farm, the greater area of influence a farm may be expected to have. Similarly, one would expect that farms that extend over a larger area should have a larger area of influence. The exact relation between farm area, weight and 'best' bandwidth choice depends on what criteria are used to determine 'best'. One simple relation is the following: for sample farm i , a good choice of bandwidth h_i satisfies the relation:

$$(4) \quad \pi h_i^2 = \lambda w_i a_i$$

where w_i is the survey weight, a_i is the farm area and λ is a smoothing parameter. The intuitive justification behind this relation is that, whereas the sampled farm may represent w_i farms, the area of farm operation is representative of the agricultural area covered by w_i farms. The degree of smoothing is controlled by the parameter λ ; large values of λ produce less detailed maps than small values. A further justification for how the survey weights and farm area should affect the area of influence may be found in Neeman (2000).

Figures 3 and 4 illustrate how this bandwidth choice solves some of the problems encountered when choosing a single bandwidth for all farms. In figure 3, the parameter λ is set to be 4, whereas in figure 4 the parameter λ takes the value 25. In both maps, the agricul-

Figure 3: Broadacre farms, 1989-99: Area operated – all

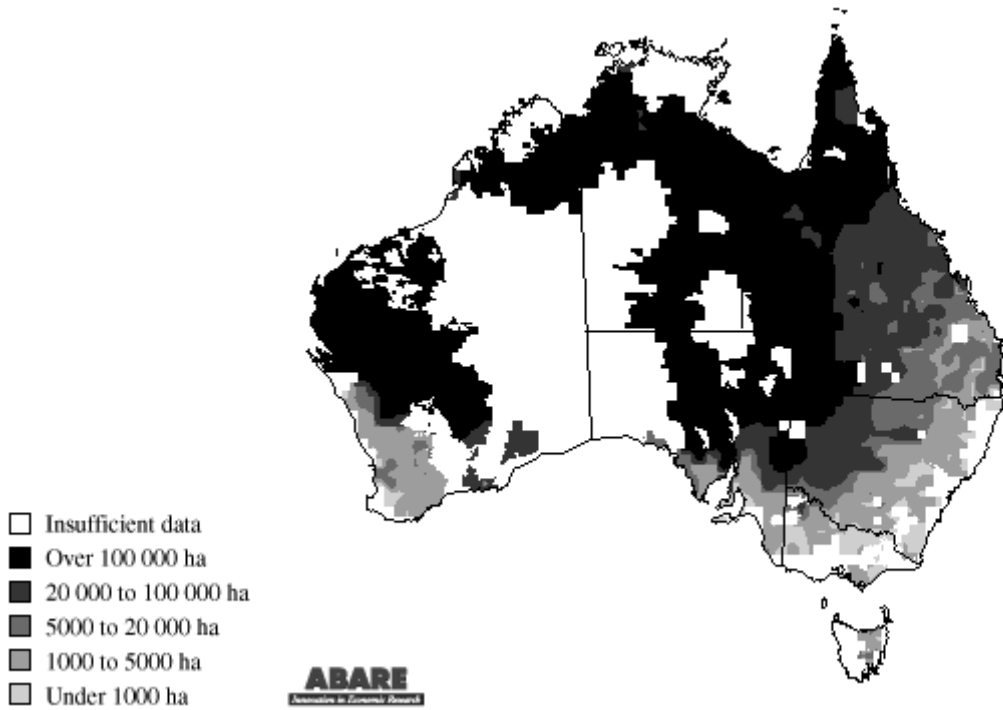
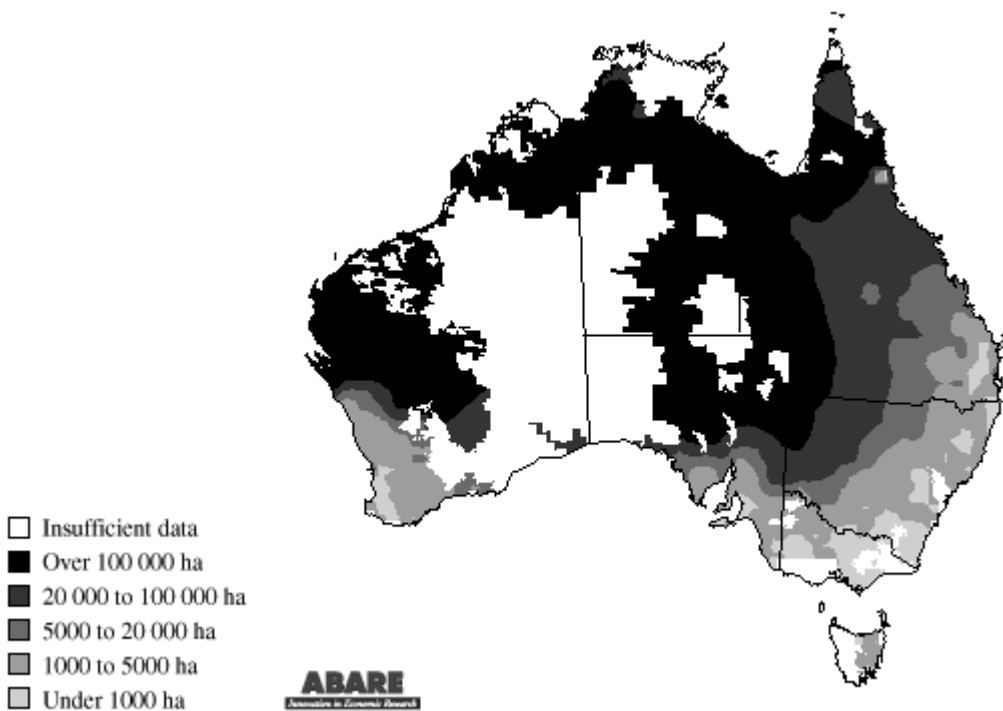


Figure 4: Broadacre farms, 1989-99: Area operated – all



tural area in the Northern Territory is well covered, reflecting the high proportion of the area that is sampled. Both maps also reflect the heterogeneity seen in the farming regions in south eastern Australia, although in figure 3, there are larger areas where data are missing.

These missing areas do not necessarily imply that there are no sampled farms in the local area. Rather, the absence of apparent information is an artifact of two major limitations in generating these maps: discreteness and masking.

Discreteness

Although in theory a local average can be generated at every pixel on the map, in practice local averages are estimated on a matrix of grid points approximately 40 kilometres apart. If a farm's area of influence does not extend to include a grid point, then that farm will not contribute to the local average at any point. Two solutions to this discreteness problem are to refine the grid matrix, or to set a minimum bandwidth to ensure that every sampled farm, no matter how small, contributes to a local average somewhere nearby. Both solutions have been explored in other research.

Masking

Masking occurs when a local average at a grid point is suppressed in order to preserve confidentiality. Depending on the density of the sample and the area of influence of each farm, local averages may be based on as few as one farm or as many as several dozen farms. When there is only one farm contributing to a grid point, then that farm's value can be determined, compromising confidentiality of the individual farm data. Once at least two farms are averaged at a grid point, it is much more difficult to uncover individual values — but then reliability becomes an issue. How well a small set of sample farms estimate a local area population average depends to a large extent on the diversity of industries and farm sizes in the area.

In figure 3, there may not be enough overlap in the sampled farms' areas of influence in the agricultural regions where small farms predominate. The same area in figure 4 has better coverage, and yet shows up much of the detail that was lost in figure 2.

While not apparent from the above maps, it can also happen that a farm's area of influence can be unrealistically large. This can happen when both the survey weight and the farm area are large. This is an unusual occurrence, since large farms generally carry low survey weights. The consequence of a large area of influence is not usually noted in the vicinity of the farm. Ironically, the farm's values can have a large influence in low density areas far away from the farm, where the sample is low. Although the farm's weight, $w_i K_h(x, x_i)$, decreases as the distance from the farm increases, the farm's weight at a

distant grid point could be large relative to the weights of other contributing farms. The easy solution to this problem is to set a maximum bandwidth. It can be shown that, within a certain range, the resulting map is not overly sensitive to the choice of a maximum bandwidth. Maximum bandwidths set in the range 400–600 kilometres produce similar results.

Conclusions

The challenges of mapping survey data require the skills of more than the statistician or cartographer. It is paramount that one has a good understanding of the data. Making the best use of survey weights in the smoothing algorithm may help ensure that the resulting maps reflect population averages both at the local and regional levels. The more one understands the geography and topography of the different agricultural regions, the better one can choose appropriate smoothing parameters. Further refinements could be made by adjusting the shape of the kernel function K_h from a circular support to a support that may depend on auxiliary information about the region or the sampled farm. One cannot ignore the technical challenges of smoothing, which include the limitations of the discrete grid matrix and the restrictions on bandwidth size. But the ultimate challenge is understanding when the map is providing the ‘right’ information; that is, information that is both reliable and easily interpretable. One must do, at some level, some ground truthing, having maps reviewed by experts in the field.

Statistical approaches, such as generalised cross-validation may be used, where the choice of parameters is based on some ‘best fit’ criteria. These criteria may vary according to the situation, and different criteria result in a different choice of ‘optimal’ parameters. For these data, for example, one could consider using a weighted least squares criterion, where the weights depended on the survey weights, on the farm areas, or on some combination of both. In the end, a well produced map of farm survey data can be a powerful visual tool for exploring data, comparing regions and generating further research in rural industry.

References

- Bardsley, P. and Chambers, R.L. 1984, ‘Multipurpose estimation from unbalanced samples’, *Applied Statistics*, vol. 33, no. 3, pp. 290–9.
- Neeman, T. 2000, ‘Nonparametric smoothing of survey data’, *Proceedings, ICES II Conference*, June, Buffalo, N.Y.