

GAPS IN MAPS, ESTIMATION OF MISSING DATA IN AGRICULTURAL STATISTICS MAPS ^α

By René Gommès¹ and Peter Hoefsloot²

1. Introduction

Starting in the early nineties, the FAO Agrometeorology Group has assembled a set of sub-national³ agricultural data in Africa - known as AGDAT - to be used as background information for crop monitoring and forecasting. AGDAT is a static database for which the years 1986-1990 were adopted as the base period (FAO, 1998).

Next to environmental information and population statistics, AGDAT contains data on 10 crops (maize, sorghum, millet, rice, wheat, common bean, cassava, taro and yams, white potatoes, cotton) and 3 groups of crops (coarse grains, roots and tubers, and pulses). For each of the crops and groups, 6 statistics are given by first sub-national administrative unit: area, production, yield, area and production per capita, and the “relative area”, i.e. the area cropped divided by the total area of the administrative unit. The basic parameters are area harvested and production by crop, but only about 60 to 70% are actually available from primary sources.

This paper presents a set of techniques which were adopted to estimate missing values of agricultural statistical data (mainly areas and productions), thus to fill the gaps in the continent-wide maps.

The basic methodology was to use pixel-based information that is generally available from non-national sources, like satellite imagery, or topographic information from digital terrain models, etc., to assist with the estimation of the missing statistical information by administrative unit (AU).

First level AUs are not necessarily the most appropriate choice for meaningful presentation of agricultural statistics, because the criteria adopted by the countries to define such units are very rarely based on agro-ecological zones. In addition, their size varies over orders of magnitude and some are so heterogeneous that the climate may vary from desert to tropical humid. AUs are nevertheless the unit used by national statistical services to report most data.

¹ Senior agrometeorologist, SDRN, FAO

² Consultant in software and analysis applied to geographical information

³ In this context, “sub-national” refers to the first sub-national administrative unit (AU) level. The actual names of the administrative units include *province, region, district, préfecture, wilaya, woreda* etc. To avoid confusion we retain the generic term of ‘Administrative Unit’ (AU) to designate all sub-national divisions

2. “Missing” statistical data

National agricultural statistics data are easily available, as they are published by the countries as statistical yearbooks and systematically assembled by FAO into the annual Production Yearbooks (FAOSTAT⁴). Sub-national data, on the other hand, are usually difficult to retrieve, and subject to a number of complications when they have to be consolidated into a homogeneous and consistent set at the continental level.

Several difficulties are associated with language, units and national typologies; they can be dealt with rather easily, and only rarely contribute to data being “missing” in a strict sense. Some examples are given below:

- “language” (*Guinea corn* is actually sorghum, *makopa* is dried cassava chops...)
- units (“bags per acre”, where the weight of a “bag” is variable from crop to crop and from country to country);
- typologies (*walo* and *dieri* in Mauritania, which refer to the type of water control...).

It occurs rather often that the data are unavailable altogether. Possible causes include:

- no sampling is carried out at the national level. Sometimes subjective estimates are produced, but they are of uncertain quality;
- data are collected at the national level, but never documented or actually published in national statistical yearbooks. However, some of those data are available nationally from the concerned services;
- different data are collected for different geographic units (for instance, not all AUs collect data for all crops, or sometimes the different AUs apply for agriculture and, say population);
- data are aggregated (by areas or by crops) in a way which is not compatible with the reporting of other countries. A typical example would be “millet” which can be bulrush millet or finger millet, or both together, although the crops are rather different from an auto-ecological point of view. The worst example being “millet and sorghum” reported as either “millet” or “sorghum”;
- data are not available for all the years from 1986 to 1990 in the reference period, or they are available only for years outside the reference period
- the AU units were modified during the reference period. This creates a minor difficulty when AU are aggregated, but when they are split, it is not always possible to redistribute the statistics between the new AU;
- the amounts cultivated and harvested are deemed to be negligible if they are below a cut-off that varies according to countries;
- crops are not reported on separately, for instance white potatoes can sometimes be lumped with vegetables and appear nowhere in the statistics.

⁴ The FAOSTAT data covering the years 1986 to 1990 are published in N. 40 to 44 of the FAO statistics Series. The information is now also available on diskettes, CD-ROM and on the FAO WWW site.

For a continental study, these practices lead to a high percentage of “missing” data. In order to fill the gaps, a number of items had thus to be interpolated or otherwise estimated.

3. Spatial interpolation of missing agricultural statistics

3.1. Overview of method

The main statistics of interest are crop production (by commodity) and the cultivated areas. Both, however, are *quantities* which are very dependent on the actual size of the AUs. It is, therefore, necessary to obtain other variables which express *intensities* like yield (i.e. production divided by area) in order to estimate them based on available information from surrounding AUs.

The following *intensities* were available:

- Area cultivated per capita (hectare/person, H/P)
- Relative area (hectare cropped/total area of AU, H/A)
- Yield (kg/hectare, K/H)
- Per capita production (kg/hectare, K/P)

The general method was thus to estimate some intensities by regression or spatial interpolation, and subsequently convert the intensities back to quantities.

Because of the practice of some countries not to report amounts (production and areas) under a given low threshold, it has sometimes been difficult to understand whether data actually were missing or whether the crop was not cultivated in the administrative unit (AU) under consideration. In these cases, agro-ecological and agro-economic conditions have been used to conjecture if the crop was cultivated or not. In most cases, however, only minor crops are left out from statistics.

The final step in the estimation procedure was a harmonisation with national annual data published in the FAO *FAOSTAT* series. This was done as follows: if Q_A is the quantity (production, area) obtained as the sum of all AGDAT sub-units, and Q_F is the corresponding value given in FAOSTAT, all AGDAT sub-national quantities (productions, areas) were divided Q_A/Q_F . The intensities were then re-computed.

Difficulties were also encountered with units and inconsistencies between individual and accumulated data items (e.g. individual coarse grains and total coarse grain values). Such inconsistencies could generally be resolved by referring to other data sources, for instance FAOSTAT.

Needless to say, this is not very simple in practice, and subjective choices had to be made.

3.2. Methods of estimation

3.2.1. Disaggregation of quantities

Disaggregation consists in re-distributing the total national production or areas in proportion to some other known variable, for instance population. This can be applied only if the country is relatively homogeneous from an agroecological point of view.

3.2.2. Estimation of intensities

As indicated above, four types of intensities can be spatially interpolated in the current context.

The most straightforward method of interpolation consists in assuming that, within a well defined agro-ecological zone, the variables above are reasonably constant.

The two first

- Area cultivated per capita (hectare/person, H/P)
- Relative area (hectare cropped/total area of AU, H/A)

are expressed relative to the known quantities of population by AU and the area of the AU. Both intensities depend on socio-economic and cultural factors and not so directly on the environmental context. Therefore, they can only be interpolated spatially, assuming that they do not vary abruptly over space. Once interpolated, they can be multiplied by the population and the area of the AU, respectively, to provide two estimates of the cultivated area.

The next intensity

- Per capita production (kg/hectare, K/P)

depends on the area cultivated, which has been obtained above. Similar to the two previous intensities, the per capita production does not depend enough on environmental conditions and can only be interpolated spatially.

The last intensity,

- Yield (kg /hectare, K/H)

depends very much on environmental conditions. It can be interpolated using both spatial interpolation and the existing links with such external variables such as vegetation indices, potential biomass and elevation.

3.2.3. Reconciling different estimates

The discussion indicates that several methods can actually be used to obtain the same quantities and intensities.

Parameter at AU level	Obtained from...
Production	1 Disaggregation of national production
	2 Spatialization of per capita production x AU population
	3 AU yield x AU cropped area
Yield	1 Regression against environmental variable
	2 Inverse-distance spatialization
	3 SEDI interpolation
Cropped area	1 Disaggregation of national cropped area
	2 Spatialization of area per capita x AU population
	3 Spatialization of relative area x AU total area

The possible data flow is illustrated in the table above. Considering, for instance, that AU yield can be derived using three methods and that cropped area can also be derived with different approaches, the number of “final” values is thus rather high. Also consider that sums can be estimated as sums (“coarse grains”) and as the individual components (i.e. maize, barley, millet, sorghum etc.), which in turn add another set of possible values. Whenever the estimates are reasonably close, they can be averages without much afterthought... but when they differ by an order of magnitude, subjective choices have to be made.

Therefore, the results are user-dependent and they result from a set of common-sense rules rather than from a method as such. It is assumed that the methods used as well as the harmonisation with FAOSTAT ensures that the final product is consistent.

3.3. Techniques of spatial interpolation

As indicated, the spatial interpolation can either be purely geo-statistical, or take advantage of the additional knowledge obtained from external variables. In the first category, the method known as “inverse distance weighting”. In the second, the method was Satellite Enhanced Data Interpolation (SEDI).

3.3.1. External variables usable for spatial interpolation

In semi-arid areas, good correlations can be found between environmental conditions and yield (K/H). Once average AU values of NDVI, elevation, etc. are available for surrounding areas, yields can be **regressed against the external environmental variables**⁵. The method is not applicable if cultivars vary in the same agroclimatic area. Generally this information is not included in the statistics, and the database was thus considered cultivar- independent. It was also not feasible to distinguish between irrigated and rainfed crops, subsistence farming and large scale modern agricultural production.

The main external variable of interest was NDVI (Normalized Difference Vegetation Index), created by NASA/GIMMS and regularly available every 10 days since 1981. It represents one of the most popular remotely sensed indicators for monitoring the response of vegetation to weather condition in several parts of Africa.

NDVI is an indicator of the density of living green mass; in theory, it varies from -1 to +1, but in practice only values between 0 to 0.7 are found on land areas.

1981-1991 average monthly NDVI data from ARTEMIS (FAO, 1993) were used to derive the NDVI variables included in the AGDAT database. The most relevant, in the current context, are NDVI monthly average, maximum and minimum together with the same values **relative**⁶ to the value of 0.12. This threshold corresponds to the occurrence of green vegetation on the ground. The interpretation of NDVI in humid tropical regions is difficult due to the absorption of infra-red light by water vapour and because the response of the index as a function of biomass reaches a plateau (saturation) at high biomass values.

Figure 1 below illustrates a typical relation between yield and NDVI, showing, among others, NDVI values starting at about 0.07 and yields levelling off from values above 0.2. Note that the values indicated correspond to the spatial average over a whole AU, where no crops are actually grown below about 0.12. This explains why relatively high yields can be found at low NDVI when some crops are irrigated or grown in the wettest parts of the AU only.

⁵ Other regression equations could be implemented, for instance those relating to population and the level of mechanisation to areas planted, etc. They were not used.

⁶ This means monthly average, maximum and minimum only for those areas where NDVI exceeds 0.12, leaving out unvegetated areas i.e. presumably areas without agriculture.

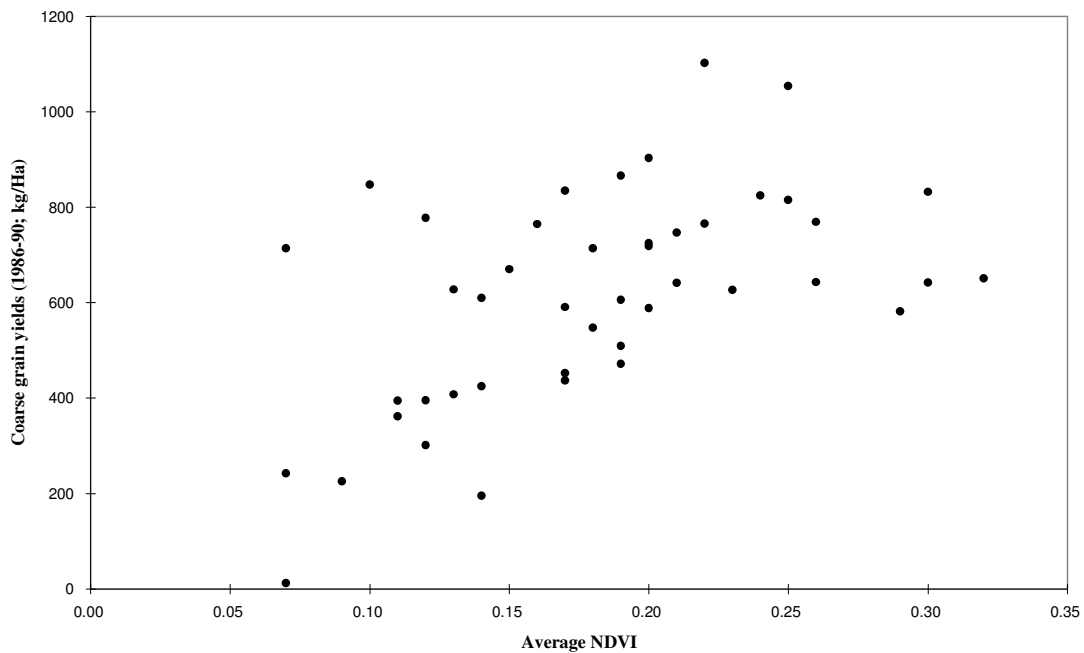


Figure 1: Coarse grain yield and NDVI in Burkina Faso and the countries surrounding Burkina Faso. The figure was restricted to yields below 1200 Kg/Ha.

3.3.2. Inverse distance weighting⁷

Geostatistical interpolation of missing data consists in the estimation of missing values at one point in space based on the known values of neighbouring entities. The *inverse-distance weighting* is one of the most straightforward methods; it takes into account the distance between the “known” and “unknown” points and their relative importance in the estimation. For instance, close-by AUs of the same country are assigned a higher weight than the AUs of neighbouring countries distance. The reason for this is that it is considered that production and feeding behaviour are more homogeneous within the country. For the AGDAT database, after testing, conditions for interpolation are at least three to ten neighbouring AUs and maximum distance between the AUs centre is 600 km.

For the geo-statistical interpolation, it was generally considered that the reported crop yield corresponds to the centre of gravity of the AU⁸. The software used for the inverse-distance weighting was mostly FAOMET (Gommes and See, 1993). Inverse-distance weighting was applied mainly to Area cultivated per capita, per capita production and relative area.

3.3.3. Satellite Enhanced Data Interpolation, or SEDI

⁷ The diskette by Bogaert et al. (1995) also includes inverse distance software.

⁸ In practice, this is the average latitude and average longitude determined from the polygons which delineate the AU.

SEDI takes advantage of the correlation between an environmental variable, for instance the above mentioned NDVI/biomass and agricultural yields. One of the ways to approach this is co-kriging, a variant of kriging using one or more auxiliary variable and exploiting both the spatial features of the variable to be interpolated and the correlations between the variable and the auxiliary variables (Bogaert et al, 1995).

The SEDI interpolation method originated in a Harare based FAO Regional Remote Sensing Project. It was originally developed to interpolate rainfall data collected at station level using the additional information provided by METEOSAT cold cloud duration images. The methods proved powerful and versatile, and it is now regularly used by FAO to spatially interpolate other parameters as well (e.g. potential evapotranspiration, crop yields, actual crop evapotranspiration estimates, etc.).

The concepts of this interpolation method and software implementing the technique have been described by Hoefsloot, 1996. The SEDI functions were recently incorporated into the WINDISP_3 software (Pfirman and Hogue, 1998)

SEDI is a simple and straightforward method for 'assisted' interpolation. The method can be applied to any parameter of which the values are available for a number of geographical locations, as long as a 'background' field is available that has a negative or positive relation to the parameter that needs to be interpolated.

Three requirements are a prerequisite for the successful application of the SEDI method:

1. The availability of the parameter to interpolate as *point data* at different geographical locations (e.g. rainfall, potential evapotranspiration, crop yields). In the present case of statistical variables, they were assigned a co-ordinate corresponding to the centre of gravity of the AU;
2. The availability of a background parameter in the form of a *regularly spaced grid* (or field) for the same geographical area (e.g. the above-mentioned NDVI variables, altitude).
3. A relation between the two parameters (*negative or positive*; Yield/NDVI is positive, temperature/altitude would be negative). A Spearman rank correlation test can reveal whether a relation exists, and how strong this relation is.

The SEDI method yields the parameter mentioned under point 1 as a field (i.e. an image covering the whole area under consideration). The average of the field value over the AU provides the estimation of the spatialised statistic. The method is illustrated below using rainfall and “Cold Cloud Duration”.

Rainfall is reported on a dekadal (10-day) basis in most countries of the world for agrometeorological purposes. A typical longitude-latitude plot showing the locations of the stations and the rainfall amounts is given in figure 2 below.

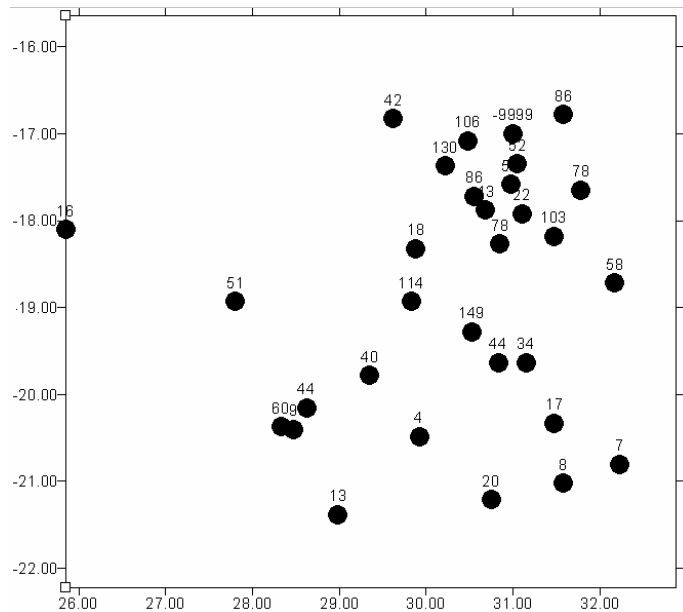


Figure 2. Rainfall (mm) for Zimbabwe second dekad of January 1991

The geostationary METEOSAT satellite takes InfraRed temperature “pictures” of the earth every half hour. In tropical regions it can be assumed that areas with temperatures lower than about minus 40°C correspond to convective cloud systems of the type that produces rainfall. The accumulated number of hours in a dekad below this low temperature threshold is known as 'Cold Cloud Duration' (CCD). It can be represented as an “image”, i.e. regularly spaced rows and columns the intersection of which corresponds to the picture elements, or 'pixels'. A pixel represents one data value⁹. Pixels can be assigned a colour depending on the value they represent, as represented by a grey-scale in figure 3.

⁹ It should be mentioned that the interpolation of point values with pixels, which represent an area of about 50 km² in the case of ARTEMIS NDVI and CCD, does entail some assumptions which are not so simple from a methodological point of view (Gommes, 1993 and 1996).

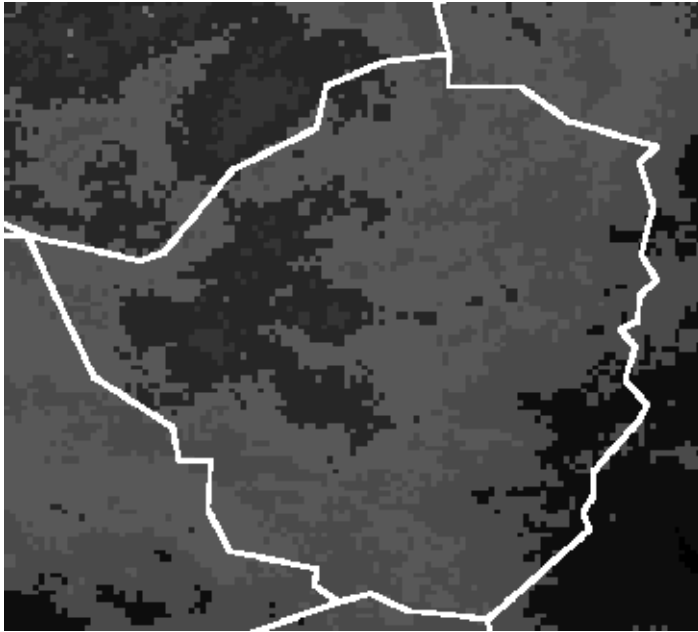


Figure 3. Cold Cloud Duration image for Zimbabwe, second dekad of January 1991

The relation between rainfall and CCD is a positive one. In other words: high rainfall values generally coincide with high CCD values.

The SEDI process is done in three steps:

1. Extracting values from the image and calculating the ratio of point and image values;
2. Gridding the ratios to form a regularly spaced grid;
3. Multiplying Grid with image to obtain estimated image.

3.3.3.1. Step 1 : Extracting values from the image and calculating the ratios

For every point value in the input rainfall data, a value can be extracted from the CCD image. The SEDI method will find the pixel that coincides with a rainfall station and extract the pixel value. In some cases the value of one pixel does not give satisfactory results. Therefore the SEDI software allows the user to extract the values of more than one pixel from the image, and take its average as image value for the station (figure 4).

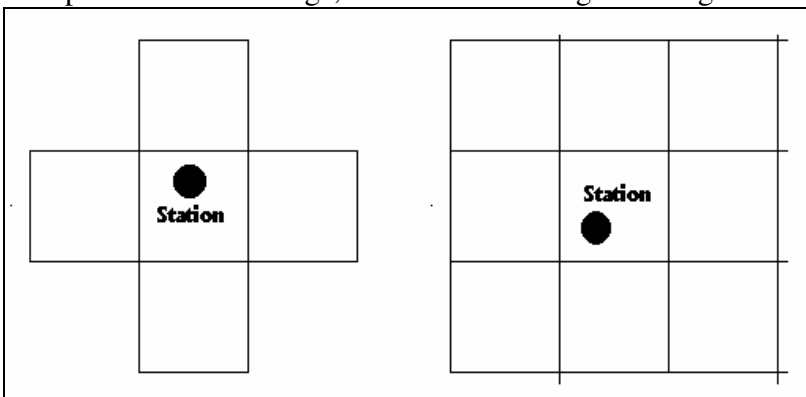


Figure 4. Extracting 5 or 9 pixels per point value

For every station we now have a rainfall value and a CCD value. The Spearman rank correlation coefficient (using the rainfall/CCD data pairs) yields a positive value. This means the relation between rainfall and CCD is positive (as to be expected). The ratio between rainfall and CCD value is now calculated as shown in table 1.

Table 1 : ratio between Rainfall and CCD (mm/hour)

Station Name	Rainfall (mm)	CCD value (hours)	Ratio
Station 1	23.4	56	0.42
Station 2	12.4	12	1.03
Station 3	54.3	96	0.57
Station 4	6.7	8	0.84

Should the relation have been negative, the ratio is calculated as follows:

$$\frac{\text{StationValue}}{\text{HighestPossiblePixelValue} - \text{PixelValue}}$$

3.3.3.2. Step 2: Creating a regularly spaced grid from the ratios

The second step constitutes of the creation of a grid from the irregularly spaced ratios, i.e. the spatial interpolation of the ratios at regularly intervals constituting a grid (figure 5).

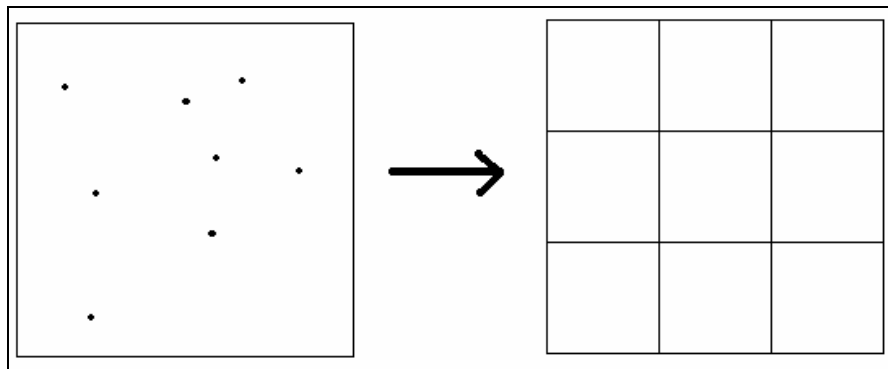


Figure 5. Creation of ratio grid from point values

The ratio grid is created with the inverse distance method mentioned under 3.3.2 with a weighting power of 2. The software allows the user to set:

- **The distance between the grid lines.** A low distance creates an accurate, dense grid, while a high value creates a coarse, less accurate and more general grid;
- **The number of stations per grid-point** determines the number of stations included in the calculation of a point in the grid matrix;
- **The maximum radius for interpolation** determines whether a value is calculated for a point in the grid matrix. If the number of stations around this grid-point within this

radius is higher than the specified number of stations, a value is calculated. Otherwise the grid-point is assigned a missing value, and the resulting image will be 'empty' at that particular point.

3.3.3.3. Step 3: Creating the SEDI image

The last step encompasses the creation of the SEDI image. The process is simple. By multiplying the grid (step 2) with the background image, an estimate for the value to interpolate is obtained. In terms of rainfall and CCD: a rainfall image is obtained by multiplying the ratio grid with the background image (figure 6).

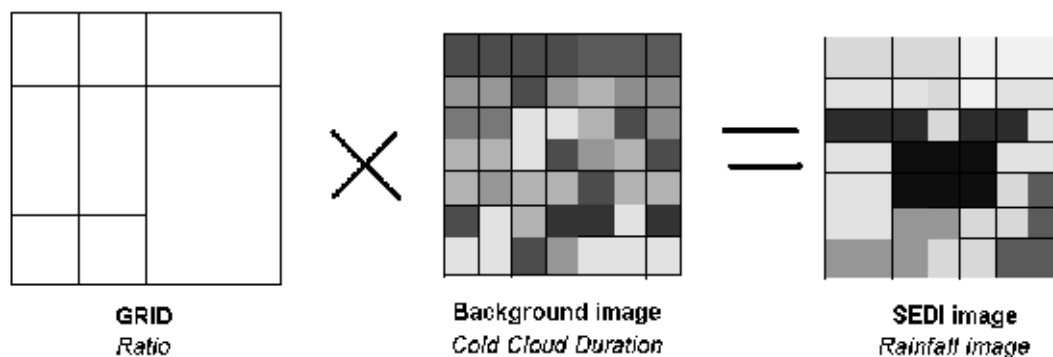


Figure 6 Creation of a SEDI image from a ratio grid and a background image

4. Conclusions

There appears to be no standard method to fill gaps in maps of agricultural statistics.

Given the nature of the underlying data, it is not even possible, in many cases, to decide that the data are actually *missing* in the sense that the crop is grown in the administrative unit under consideration, but no data are being published nor are they available from informal national sources.

The methods of estimation apply at the continental scale; they all take advantage of the knowledge of the same statistic from neighbouring areas, as well as of the statistical links between “global” environmental variables and the agricultural statistic. Most global environmental variables which have been used are derived from environmental satellite observations.

One of the main problems, in fact, is that several methods can be applied to derive the same parameter, say agricultural production, and that it is not known which method will provide the most reliable result. A common-sense decision has thus to be made between the various options, particularly if they diverge markedly.

The final step is an empirical reconciliation with published national statistics. This ensures that no gross errors are introduced by the interpolation, even if some smoothing is no doubt the result of the procedure.

The above all points at the fact that filling gaps in maps is art at least as much as science.

5. References

Bogaert, P., P. Mahau and F. Beckers, 1995. The spatial interpolation of agroclimatic data. Co-kriging software and source data. User's manual. FAO agrometeorology working paper series N. 12. FAO, Rome. 70 pp + 1 diskette. The programme is retrievable from <FTP://FTP.FAO.ORG/SDRN/Co-Krig>.

FAO, 1993. FAO-ARTEMIS NOAA AVHRR NDVI Image Bank, Africa 1981-1991. ISO 9660 CD-ROM. FAO, Rome.

FAO, 1998. AGDAT, Sub-national agrometeorological database for Africa. FAO agrometeorology working paper series N. 14. FAO, Rome. 65 pp + 1 diskette. IN preparation.

Gommes R. and See L. 1993. FAOMET. Agrometeorological crop forecasting tools. FAO agrometeorology working paper series N. 8. FAO, Rome. 57 pp + 1 diskette.

Gommes, R. 1993. The integration of remote sensing and agrometeorology in FAO. *Adv. Remote Sensing*, 2(2):133-140.

Gommes, R., 1996. Crops, weather and satellites: interfacing in the jungle. Proc. of COST 77 workshop on **The Use of Remote Sensing Techniques in Agricultural Meteorology Practice**. Budapest, 19-20 Sept. 1995. Pp. 89-96 of EUR 16924 EN, EU Directorate-General Science, research and Development, Luxembourg: Office for Official Publications of the European Communities, 289 pages.

Hoefsloot, P. 1996. IGT manual, Ver. 1.10. Working paper series N. 5. SADC/FAO, GCPS/RAF/296/NET, Harare, 53 pp. Programme and manual are retrievable from <FTP://FTP.FAO.ORG/SDRN/IGT>.

Pfirman, E., and J. Hogue, 1998. WINDISP_3. Programme and manual are retrievable from <FTP://FTP.FAO.ORG/SDRN/WINDISP3>.

^α Gommes, R., and P. Hoefsloot, 1998. Gaps in maps, estimation of missing data in agricultural statistics maps. Pp 155-168 in : Proc. of the the EU/COST-79 Seminar on Data Spatial Distribution in Meteorology and Climatology, Volterra, 28 Sep.-3 Oct. 1997), EU, Luxembourg, EUR18472, 226 pp.