Africa Soil Information Service

SPATIALLY EXPLICIT AND EVIDENCE BASED

SOIL MANAGEMENT RECOMMENDATIONS

TECHNICAL SPECIFICATIONS

DATA QUALITY CONTROL PROCEDURES FOR DATA PERTAINING TO THE DIAGNOSTIC TRIALS

E. Jeroen Huising, Job Kihara and Generose Nziguheba

DATA QUALITY CONTROL

This section deals with the quality assurance of data pertaining to the diagnostic trials with special reference to the crop response (harvest) data recorded in the field. Data quality is generally defined as the fitness for use of a specific data set. In this case it refers to the use of the data for analyses of crop response to the various treatments and variability of the crop response within and between sentinel sites. This requires a certain level of accuracy and consistency in the way the data is collected, handled and presented for each individual sentinel site.

AfSIS has a standard protocol for the implementation of the diagnostic trials and standard forms for registration of field observations (the log book). An ideal situation to ensure data quality is that the same team conducts the diagnostic trials, does the data collection and recording to assure consistency in the way the trials are implemented and data is recorded. Nevertheless, different teams operate the diagnostic trials at the various sentinel sites. Training of the team members on the protocols and procedures is an important aspect of quality assurance. Even with trianing, making mistakes is human and there are various sources of error that may affect the quality of the data. Quality control procedures are therefore required to check on the consistency and accuracy of the data that will then allow for integrated and comparative analyses of crop response for individual trial sites and sentinel sites. For understanding of the quality control procedures described in this section we assume familiarity with the protocol for the diagnostic trials.

This section explains the data quality control procedures and criteria used for quality control as well as procedures for correcting the data in case these contain (or is perceived to contain) errors. Data quality control is an important aspect of data management, handling and processing. Errors easily propagate through the various processing steps, affecting the reliability of the outcome and therefore the validity of the conclusions that are drawn based on the results from the diagnostic trials.

DATA QUALITY ASPECTS AND SOURCES OF ERROR

Data quality is defined by the following components: lineage, accuracy (both positional and attribute accuracy), logical consistency (or referential integrity) and completeness.

Lineage is concerned with the historical and compilation aspects of the data, such as:

- Source of the data: In our specific case, who in which national team has done the recording? Also, AFSIS will increasingly use data from diagnostic trials that have not been implemented by AfSIS itself but by projects and initiatives that are associated with it, and where AfSIS has less control over how the trials have been implemented, and for which we may not have specific details.

- Content of the data: We may not always know what exactly has been recorded. There may have been additional treatments or additional trials. There may be data on site characteristics, rainfall data etc. It is important that the content is specified.
- Data capture specifications: It is often not specified what exactly has been captured. For example, in our case we measure the weight of the cobs harvested in the net plot, while at the same time we select a subsample of the maize cobs for determining the dry weight percentage. Whether the weight of the cobs in the subsample is included in the recorded weight of the harvested cobs depends on the procedure followed. Then we have to verify whether the subsample actually consisted of 5 cobs as specified in the protocol. The netplot that is harvested is specified by dimensions as indicated in the protocol, but also by the number of rows and plants per rows in the net plot. What if the plant spacing was not according to the specifications, are the dimension adhered to during harvesting or is the harvesting done according to the number of rows and plants per row that is supposed to constitute the netplot. Similarly, how is the height and basal diameter of the plant measured? Data capture specifications also refers to the instruments used in the recording (see under accuracy and precision).
- Compilation method: Is this done by hand or electronic? Also are various sheets used and is the data compiled onto one sheet later? What have been the different steps in the data compilation?
- Transformation and data conversion: at what stage is the conversion of grain yield to kg/ha or t/ha, for example, done? Has it been done correctly

Accuracy: For accuracy assessment we have to consider the inherent and operational error. The inherent error is associated with the instrumentation. We use electronic balances, spring balances and other devices. These instruments have a certain precision and if not very precise these will add to the variability in the measurement and therefore error. E.g. we have observed measurement of weight of stover or cobs that are recorded in steps of 0.5 kg, for specific fields rather than all fields in the sentinel site. Likewise, instruments need to be calibrated and some balances are notoriously unstable. If not properly calibrated, this will cause systematic errors that affect the accuracy of the measurement.

Logical consistency or referential integrity: this generally refers to relational databases, but is equally valid in our case. This relates, for example, to whether the reference to the field (i.e. one trial) is done consistently. The field ID is indicated in the spreadsheet where the plant growth characteristics are recorded, in the sheet where the plot data is recorded and in the sheet where the field data is recorded. They all have to refer to the same field. We have seen fields with duplicated farmer names or records that are identified by the farmer name where these do not appear in the field data. Likewise, the location of the field may have been changed and there are multiple records of the location of a particular field. The latter relates to referential integrity.

Completeness: Generally data sets are not complete, in the sense that harvest of the plot may have been lost due to adverse weather conditions, or because the farmer harvested the plot or other. You also find that the data was simply not recorded or could not be established (e.g. dates of weeding or other field operatios). This is different from holes in the data as result of data having been eliminated because of the quality control procedures. In all cases this needs to be documented. In most cases corrections are made to the data as part of the quality control. It is important that these remain visible in the original data files by either highlighting the particular data field or using a different font colour and by inserting comments documenting the corrections or change made.

PROCESS OF DATA QUALITY CONTROL

Data quality control is not one particular step in the data handling and processing. Quality check should be carried out at the different stages of the process, from the recording in the field and in the lab, to transferring the data to electronic form, copying data, basic data conversions, and calculations carried out on the data. It is best to, as much as possible, standardize the processes, keep track of the data files, standardize naming conventions of the files, and keep backups. It is always a good idea to involve as few people as possible in this process and for specific tasks and restrict access to data files to unauthorized persons and it is certainly advisable to make one person responsible for data quality aspects through the whole process.

You cannot avoid making mistakes; in all cases above, humans do the recording and humans make mistakes, either in the reading of the instruments or copying the values to either paper or entering it into the computer. Errors may occur in the calibration of the instruments, in the labeling of the bags in which the samples are kept and there may be effects of the storage of the samples and may even result in loss of samples all together, to name but a few.

Despite that much of the potential errors are influenced by the particular circumstances, it is important to record and document these as much as possible. Plants may have lodged, been eaten by cattle or attacked by termites. Birds might also have eaten from the cobs or there may be otherwise damage to the cobs. Additionally, cobs and/or plants may have been lost, making it difficult to make an accurate assessment of the number of plants, number of cobs and subsequent reliability of the stover and or cob weight recorded. We have tried to capture as much of these effects in the recording sheets as possible. Make full use of the opportunities on the recording sheets to make remarks.

As far as the quality control and assurance is concerned, all changes to the data sets need to be documented for future reference. We have used a form to document the various data quality aspects of the data sets, to log the various data quality control activities and changes we have made to the data set, to sign off on quality checks performed and to provide summary statistics on the data contained within the data set. The form is provided in appendix A

In this section we will deal with quality checks on the data as received in electronic forms from the responsible field and laboratory teams to assure minimum data quality standards before it is entered into the database and /or made suitable for further analyses.

DATA QUALITY CONTROL -CHECKING FOR OUTLIERS AND ERRORS IN DATA ENTRY

Harvest area:

a) Check whether the harvest area refers to the effective area occupied by the number of rows and number of plants per row that constitute the net plot, taking the plant spacing into consideration. This is dependent on the test crop used and the way the trials are implemented. E.g. if the netplot is constituted of three rows of maize spaced at 75 cm and 12 planting station per row at 25 cm spacing, the harvest area is 60*0.75*0.25 = 6.75 m2 (rather than the nominal value of 9 m^2 (i.e. 3 x 3 m) indicated in the protocol. Similarly, if the whole plot is harvested the effective area is 26.25 m^2 (i.e. 7*20*0.75*0.25) rather than 25 m^2 (our plots measure 5 x 5 m).

Plant count and number of cobs (sorghum: heads) harvested

b) Flag those entries that have less than a third of the original number of plants remaining (irrespective of whether the netplot or whole plot is harvested). Depending on the cause of the loss of plants, harvesting the remaining plants will cause a bias as the stronger plants may have remained, tending to increase the grain and stover yield if this is adjusted for size of the net plot.

c) If absolute counts drop below 10, the results obtained from that plot is generally not considered representative or very reliable.

d) If plants standing and plants lodged have been counted, verify whether the plant count refers to the sum of both, in which case it is assumed that plants have lodged but have not been lost. The number of plants and associated data on stover weight and cob weight (as well as number of cobs) is assumed to be a reflection of the true values (verify from the report on the data collection). If lodged plants are extensively damaged (e.g. termites feeding on the lodged plants and cobs) these can be left out and adjusted as missing plants.

e) Calculate the average number and standard deviation of number of cobs per plant. Flag those that are more or less than the average plus or minus two times the standard deviation. Check if there is a recording problems in those flaged records and decide whether to keep or to exclude e.g., a value might be retained if it conforms to similar pattern as other treatments for that particular field or of the same treatment in multiple fields.

Weight of cobs harvested

a) Calculate the average cob weight for each plot; that is the total fresh weight of the cobs harvested divided by the number of cobs. Calculate the average and standard deviation and flag those that are more then AVR+2*STDEV and less than AVR-2*STDEV.

b) Check and compare the average weight of cobs harvested with the average weight of the cobs in the subsample. It is assumed that a subsample of 5 cobs is taken from the total number of cobs harvested unless otherwise stated in the report (check). The subsample should be representative of the cobs harvested and should therefore be of similar average weight. Flag those entries of average cob fresh weight where the discrepancy is big (more then 2*average cob fresh weight of the subsample or less than 0.5* average cob fresh weight). (This criteria can possibly be tightened: calculate the ratio of average weight of the cobs harvested and of the cobs in the subsamples and determine average and standard deviation – apply the criteria of two times the standard deviation to flag outliers).

c) If a high average cob weight corresponds to a relatively low *CobNo*/*PlantNo* ratio or vice versa a low average cob weight corresponds to a relatively high ratio of cobs per plant, one may expect there is an error in the number of cobs recorded (i.e., the average cob weight criteria no longer apply).

d) If otherwise, consider correcting the total cob fresh weight if it is deviating from measurements for corresponding or similar treatments in that field and if corresponding total grain yield calculated is considered suspicious (see calculating the total grain yield and adjusted total grain yield). Corrections should be done using the cob number and the average cob weight for corresponding or similar treatments in the field (e.g. if it concerns a NPK or NPK + treatment, take the average cob fresh weight from all NPK and NPK+ treatments).

Cob fresh weight of the subsample, grain fresh and dry weight

a) If the grain fresh weight of the subsample (*SSGrainFW*) is provided, calculate its ratio over the fresh weight of the cobs (*SSCobFW*) and determine the average and standard deviation. Flag those entries that are more then AVR + 2*STDEV and those that are less than AVR - 2*STDEV. Flag any entries that are more than 1.00.

b) Calculate the ratio of the grain dry weight and grain fresh weight (*SSgrainDW*/*SSgrainFW*). Determine average and standard deviation and flag entries that are more then AVR + 2*STDEV and those that are less than AVR - 2*STDEV. Flag entries that are more than 1.00

c) If (*SSGrainFW*/*SSCobFW*) is within range, correct the *SSGrainDW* using the average ratio (*SSGrainDW*/*SSGrainFW*) for that field (or use the corresponding treatments within that field to calculate the average ratio). Set the dry weight to fresh weight if dry weight still higher than fresh weight after correction.

d) If the (*SSGrainDW*/*SSGrainFW*) is within range, it is most likely that the error is in the recording of the *SSCobFW*. Check and confirm with the average cob weight of the subsample with the average cob weight for the plot. Correct using the average (*SSGrainFW*/*SSCobFW*) ratio for the field, or of the corresponding treatment within the field if there is strong variation between treatments. In excel the formula would look like the one below.

$$SScobFW_{ij} = SSgrainFW_{ij} \Big/ AVERAGE_{i}\left(RRSSgrainFWcobFW_{i,j=1} : RRSSgrainFWcobFW_{i,j=10}\right)$$

where: *RRSSgrainFWcobFW* = the ratio between *SSGrainFW* and *SSCobFW*
i = is the field number
j = the plot number

e) If both (*SSGrainFW*/*SSCobFW*) and (*SSgrainDW*/*SSgrainFW*) are outside the range it should be investigated whether the probable cause is a wrong recording of the *SSGrainFW*. If so, correct using the average (*SSgrainDW*/*SSgrainFW*) ratio or the (*SSGrainFW*/*SSCobFW*) which ever gives the most likely outcome. Alternatively the average (*SSgrainDW*/*SSCobFW*) ratio can be used in the calculation of total grain yield, without necessarily having to adjust the individual values. This will however not affect the total grain yield calculation. *SSGrainFW* should be equated with the *SSCobFW* if *SSGrainFW* results to be larger than the *SSCobFW* after correction.

f) In case the moisture content of the grain is determined rather than determining the dry weight, the check can be done on the moisture content directly by determining the average and standard deviation and flagging the outliers. Similar procedures are followed as under a, b, c, d and e above. Values can be dropped if the source of error cannot be verified and the above criteria is not helpful.

DATA CONVERSION AND BASIC CALCULATION OF STOVER AND GRAIN YIELD

Total grain yield

The grain yield is specified in tons/ha. It is calculated using the general formula below:

*TGrainYld = TCobFW * (SSGrainDW/SSCobFW) * (10/Harea)*

Any correction done on the *TCobFW*, *SSGrainDW* or *SSCobFW* will automatically be reflected in the total grain yield. The above formula will provide the correct unit conversion if TCobFW is stated in kg and Harea in m^2.

Total stover yield

Similar formula is used for the calculation of the stover yield in dry weight per ha

*TStoverYld = TStoverFW * (SSStoverDW/SSStoverFW) * (10/Harea)*

Adjusted harvest area

The plant count and the row and plant spacing, using the formula below, determine the adjusted harvest area:

*AdjHarea = Plantno * Rowsp * Plantsp*

Adjusted total grain yield

We generally find a large variation in the number of plants recorded in the plot (whether in the net plot, but even more if the whole plot has been harvested) to the extent that this interferes with the results from the analyses. That is, there is a of lot variation that is observed in response of the crop as expressed by the total grain yield when comparing between treatments and between sites, which is actually explained by the variation in the number of plants harvested in the plot. We often see that this is a random effect that cannot be attributed to the particular treatment. Consider the data example below from the Kiberashi (Tanzania) sentinel site, to illustrate this point.

We therefore calculate an adjusted total grain yield (*TGrainYld_adj*) that corrects for the plant count, using the adjusted harvest area (*AdjHArea*). We likewise find a lot of variation in the number of cobs per field, but these are considered to be an effect of the interaction between the treatment and soil condition and are therefore not corrected for in the calculation of the *TGrainYld_adj*. Note that we provide both the adjusted and unadjusted yields in our datasets to suit the different analyses that may be preferred.

The general formula for calculating the *TGrainYld_adj* is similar to the one for total grain yield:

*TGrainYld_adj = TCobFW * (SSGrainDW/SSCobFW) * (10/AdjHarea)*

Corrections on the data set, considering all the data fields discussed above, are done after the *TGrainYld* and *TGrainYld_adj* are calculated. We will make an assessment on a field-by-field basis, whereby any record in which any of the above attributes are flagged as outliers are scrutinized. Apart from this we identify *TGranYld_adj* values as 'suspicious' if they are too high in absolute terms (e.g. maize yields of 12 t/ha if the yield potential is 8 t/ha) or in relative terms (e.g. we do not expect the yield of the NPK+manure treatment to be twice that of the NPK treatment). In that case, we look at the above variables (data fields) and consider their value in relation to the values for that variable in the other records pertaining to that particular field, to identify 'local' outliers. These 'local' outliers clearly deviate for the attribute values in the other records in the field data subset, without fulfilling the formal criteria for outliers as mentioned above. This assessment considers the variability within the field, though it is difficult to provide formal criteria. However, consider the following:

a) Low plant numbers tend to increase the *TGrainYld_adj* and where suspiciously high *TGrainYld_adj* values coincide with absolute low plant numbers the *TGrainYld_adj* value is excluded. If all observations regarding the cobs harvested (the numbers, weight, dry weight percentages, etc.) seem normal and trustworthy the plant number can be corrected using the cob number and inverse of the average cob per plant ratio for the field. Rather than using the average cob/plant ratio considering all treatment for that field, it is advised to consider only the treatments that give similar crop response; that is if it concerns a control treatment consider the other control treatments and the treatments that give similar poor response. Similarly if it concerns a treatment that gives good response consider the other treatments that give good response in determining the cob/plant ratio to be used (see Table 1).

b) All records that are flagged for containing outliers are not automatically excluded, or even mean that corrections are needed. This will depend on the general pattern for that particular field. If, for example, the average cobs weight is (too) high, but consistent with general very high cob weight for the other treatments the value can be retained. That is it is not considered a 'local' outlier. Yields should, however, not exceed the yield potential for that variety. The same principle applies for outliers on the low side.

c) Zero (0) values are allowed if there is no yield and if this is not expected to be the results of loss of, or feeding on, the crop, or of the crop having been harvested by the farmer or other. We leave the entry blank (Null and void) for if the data is missing or data entries excluded/rejected for any reason; we maintain the zero value for if actually zero yield was measured. At the analysis stage and depending on the objectives, one can decide to drop values below a given threshold.

d) All abnormalities in the plot management, like top dressing having been done weeks too late, or plot not having been weeded at all, automatically lead to exclusion of the entry (null and void).

e) In first instance corrections are made to the input variables (values of the individual attributes). However, if not obvious which attribute value to correct or multiple attributes need to be corrected, changes can be made to the formula directly where the average ratio for the whole field is taken as a coefficient in the equation for converting to dry weight for example. In particular cases, when the dry weight/ fresh weight ratios are very variable for that field the same coefficient can be used for all observations for that particular field.

f) In case the dry weight has not been determined, but the moisture content of the grain is measured, the conversion to dry weight is done assuming 12% moisture content for dry grains.

Adjusted total stover yield

The formula for calculating the adjusted stover yield in dry weight per ha is:

*TStoverYld_adj = TStoverFW * (SSStoverDW/SSStoverFW) * (10/AdjHarea)*

There are fewer options to check and correct for wrong entries. Because of the large variability, we consider it not applicable to check on the average weight of the plants and to compare with the average weight (dry or fresh) of the plants in the subsample. However, any change made to the *AdjHarea* as consequence of corrections made to the plant number harvested based on considerations of cob numbers in relation to the plant numbers as discussed above, will automatically be reflected in the *TStoverYld_adj* .

Table 1. Variation in plant numbers harvested in the netplot per treatment, plot and field for the Kiberashi sentinel site season 2009-2010

Fields	**plot 1**	**plot 2**	**plot 3**	**plot 4**	**plot 5**	**plot 6**	**plot 7**	**plot 8**	**plot 9**	**plot 10**	**Sum**	**Average**	**Std dev**
1	15	24	22	24	26	21	20	25	23	20	**220**	**22.0**	**3.20**
2	10	14	28	19	22	31	23	17	19	21	**204**	**20.4**	**6.19**
3	23	20	20	21	23	19	25	25	28	28	**232**	**23.2**	**3.26**
4	20	26	25	16	22	22	20	22	23	19	**215**	**21.5**	**2.92**
5	23	20	19	11	12	25	20	15	19	16	**180**	**18.0**	**4.50**
6	11	10	14	10	14	25	14	12	9	17	**136**	**13.6**	**4.70**
7	12	18	22	11	18	15	15	19	12	30	**172**	**17.2**	**5.71**
8	25	28	28	29	30	23	28	30	27	25	**273**	**27.3**	**2.31**
9	11	12	15	12	16	16	24	18	17	12	**153**	**15.3**	**3.92**
10	23	27	25	23	27	21	22	23	36	26	**253**	**25.3**	**4.30**
11	17	18	21	10	12	9	15	13	8	23	**146**	**14.6**	**5.10**
12	19	12	17	17	16	14	14	10	11	12	**142**	**14.2**	**2.97**
13	23	25	23	29	22	19	20	20	27	25	**233**	**23.3**	**3.23**
14	21	25	29	20	25	22	17	19	21	23	**222**	**22.2**	**3.46**
15	25	29	24	18	24	20	28	29	26	31	**254**	**25.4**	**4.12**
16	31	30	27	20	21	19	27	13	22	34	**244**	**24.4**	**6.47**
SUM	**309**	**338**	**359**	**290**	**330**	**321**	**332**	**310**	**328**	**362**			
Field count	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**			
Average	**19.3**	**21.1**	**22.4**	**18.1**	**20.6**	**20.1**	**20.8**	**19.4**	**20.5**	**22.6**			
Std dev	**6.12**	**6.57**	**4.60**	**6.26**	**5.43**	**5.11**	**4.85**	**6.00**	**7.72**	**6.48**			
Treatment	NPK	NP+lime	Control	NPK+Man	NP	NPK+MN	PK	NK	NPK	Control			

Above we see a data example from the Kiberashi sentinel site with little variation in soil condition and fairly good management of the diagnostic trials. We see that when averaging across the 16 fields that there are large differences between the treatments. The NPK+Manure treatment, on average, has 80% of the number of plants counted for the control treatment in plot 10 (18.1 vs. 22.6). Likewise, the plant count in field 6 (13.6) has is on average 54% of the count in field 15 (25.4). WE also observe the large standard deviation in plant count both for the various treatments and for the various fields.

CHECKING FOR CONSISTENCY AND REFERENTIAL INTEGRITY

We tend to keep data in different workbooks or in different worksheets within a workbook (if EXCELL terms are being used) and we often duplicate information in these workbooks and spreadsheets. Like in our case, the information on the cluster, the field and plot numbers, sometimes the names of the farmers and the coordinates of the location of the fields, are often duplicated. The files with the spectral data from the soil and plant tissues samples, as well as the chemical analyses of the soil and plant tissue samples also refer to the site, the cluster, field and plot. Often this data is processed by different people who may have used an internal coding system for administrative purposes and used different ways of referencing to the site, cluster and field. Moreover in most of the sites, the land degradation surveillance framework (LDSF) has been applied and we will reference to those data sets for further analyses of the data. It is therefore important to ensure consistency in the referencing

a) Check for duplication of the field and plot numbers/ids in each of the spreadsheets and data files and correct. In case we have a lot of duplication of field numbers (in the sheet that contain the plot data) and plot numbers (in the sheet that contains the plant data) check for typing errors.
b) Check for consistency between the various data sheets and data files in referencing to the cluster, field and plot. That is, the field id's that appear in the PLOT data sheet should be the same as those that appear in the FIELD data sheet. If this is not the case this should be corrected, or if records pertaining to a particular field are missing in the PLOT data sheet it should be indicated in the FIELD data sheet (where the record pertaining to that particular field may be retained).
c) Special attention needs to be devoted to the data consistency when the trials in a particular site are repeated in following seasons. If trials are repeated on exactly the same location the same identifier for that field should be used. If however the site is moved to a different location, even within the same farm, a different identifier needs to be used. If trial sites are renumbered use a link table to relate old and new trials.
d) If there are files that contain the same type of data, like for example if there are different files with the coordinate data for the location of the fields, consolidate and delete the file that has become redundant (or link the files through a table).

DOCUMENTATION

Please find appended the format for the file we use for documentation of the data files. It provides for entering administrative data like the location of the file and source of the data etc. It allows for description of the content of the data file, to maintain a log of all the work done in relation to quality control, signing of on the quality checks done and to enter summary statistics on the number of records in a data set. There should be one documentation file per data set.

If approved and the various quality control checks have been signed off, the file should be write protected, with only authorization to make changes to one (or very few) person(s). It is only then that the separate data sheets can be exported to CSV files, for further use or entering into the database. The CSV file will contain only the data values, whereas the original file will contain all the formulas. All future corrections will have to be done on the original data file, which will remain as the only reliable data source for further application.

APPENDIX A: FORM FOR DATA DOCUMENTATION AND QUALITY CONTROL ON DATA SETS PERTAINING TO THE AGRONOMIC TRIALS

General information	
Data file: name	Name of the data file [Name of the documentation file]
Data file: type	Extension of the file (file type)
Data file: location	Computer/server where the data file is kept and directory path
Source of the data	Name of project, institution and/or person involved in the data collections and that has provided the data
File history	Indicate previous versions of the data set in so far these exist as separate files or the different data sets that have been compiled into this data set that may contain the original non-modified data
Content	Describe the content of the data file and how the data is organized (that is the different worksheets)
Worksheets and data fields (columns)	
Worksheet:	Name of the worksheet for which the data fields are described below.
<Data field: name>	Give specification of the data contained in this column, including, if possible, specification on the data capture. Should capture the data from the report on the data collection for the diagnostic trials. Should certainly reflect if measurements are done differently than indicated in the protocol (e.g. number of cobs and plants in the subsample – or of best plants harvested)
<Data field: name>	Repeat for as many data fields in this worksheet
Etc.	
Worksheet:	Repeat above for as many worksheet contained in this data file
<Data field: name>	
Etc.	
Logbook	
<Date>	<Name> Name of the person who has done the quality checks and corrections
	Describe the actions taken to check the data or corrections made or calculations done; itemize: use separate entries for each type of action/activity
	Repeat above for each individual step in control and processing of the data
	Etc.
<Date>	<Name> log activities per date
	Etc.
Quality check	
☐ <Name> <Date>	**Missing values (completeness) and consistency>** Identify the missing records, the fields for which the records are missing and individual plot records that are missing; count the number of records that are

	incomplete (for which esp. the grain yield data is missing) and indicate the reason why, if that is known Sign off or tick box if task completed
□ <Name> <Date>	**Formulas and calculations**> Check on all formulas used in spreadsheet for conversion, calculation of stover and grain yields and other (e.g. whether average and standard deviation is referenced to the proper range of cells, based on which the criteria for the quality control are determined) Sign off or tick box if task completed
□ <Name> <Date>	**Outliers and corrections>** Remark on, or refer to the logbook entries, to indicate type of corrections done and the specific considerations and interpretation of the criteria used to reject, accept or correct entries in this data set (especially with reference to the calculation of the grain yields), since every data set is different Sign off or tick box if task completed
Summary statistics	
<nr. of records>	Enter the number of record contained in the file related to the agronomic trial sites (fields); <nr of incomplete records> Give nr of records that is incomplete (for mandatory data fields)
<nr. of records>	Enter number of records of plot data contained in the file; mention separately the fields for which the plot data is missing, and apart from that specify the number of individual records for which grain yield is missing
<nr. of records>	Number of records for plant height and BC measurements; Indicate if there are obvious omissions in plant height and BC recordings.
General remarks	
	General remarks and observations regarding the fitness of use for further analyses. i) There may be data, other than the field, plot and plant data, contained in this file that has not been checked, and further use should be cautioned ii) There may be general observations on the fitness for use for particular analyses, e.g. there may be a bias towards high calculated grain and stover yields (higher than expected attainable yields) that will still allow for analyses of variation between fields an treatments within the sentinel site but not between sentinel sites without further corrective measured iii) There may be observation on the variability of the individual plot level measurements that makes comparison between fields difficult, but would still allow for comparison between sentinel sites