Food Sec.
DOI 10.1007/s12571-017-0714-y

ORIGINAL PAPER

Post-harvest management and post-harvest losses of cereals in Ethiopia

H. Hengsdijk[1] · **W. J. de Boer**[1]

Received: 21 July 2016 / Accepted: 25 July 2017

Abstract Recent and systematic evidence on the magnitude of post-harvest losses in sub-Saharan Africa is scarce, hindering the identification of interventions to reduce losses. Here, we unlock standardized and systematically collected information on post-harvest management and farmer-reported post-harvest loss estimates from the Living Standards Measurement Study – Integrated Surveys in Agriculture. Using the data from Ethiopia, the objective is to disentangle factors that induce or relate to post-harvest losses in cereals. The data of approximately 2500 households and 5500 cereal records were analysed. Cereal post-harvest loss was reported by only 10% of these households. The average self-reported post-harvest loss was 24%. Rodents and other pests were most frequently reported to cause these losses. Adoption of improved storage methods was limited and most cereals were stored inside the house in bags. Random Forests (RF) was applied to gain insight into factors and conditions favouring post-harvest losses. Application of RF explained 31% of the variation in post-harvest losses reported by farmers. Three major factors associated with post-harvest losses were the distance of the household dwelling to the nearest market, the distance of the household dwelling to the main road, and average annual rainfall. Losses increased the further households were located from a market or main road, and losses also tended to decrease with higher rainfall. The standardized and nationally representative survey data from Ethiopia used were a good starting point for modelling post-harvest losses but the finally available information appeared to be partial. Therefore, this paper calls for better data collection, which could help to better target interventions needed to reduce post-harvest losses.

Keywords Food loss · Random forests · Food storage · Maize · LSMS-ISA

1 Introduction

Recent studies have highlighted the large food losses which occur after crops are harvested up to the times when they are consumed (FAO 2011; Lundqvist et al. 2008). Reducing food losses could be a major contribution to satisfying anticipated higher global food demand and to improving food security and resource use efficiency (Godfray et al. 2010; West et al. 2014; Hertel 2015; Reynolds et al. 2015). There seems to be consensus in the literature that post-harvest losses in developed countries are relatively high at the consumer end, while in developing countries they are relatively high in the early stages of the post-harvest system i.e. at farm level (Parfitt et al. 2010; Hodges et al. 2011). However, recent and systematic evidence is lacking on the magnitude of post-harvest losses at farm level in developing countries. In sub-Saharan Africa (SSA), post-harvest loss studies at farm level almost exclusively focus on storage losses (www.aphlis.net; Rembold et al. 2011). In most studies non-standardized and biased methodologies are used and estimated storage losses are generally inaccurate (Affognon et al. 2015). Consequently, there is a lack of reliable information on the post-harvest losses faced by farmers. More generally, there is a lack of information on the post-harvest management by farmers and the conditions under which they operate in relation to post-harvest food losses.

✉ H. Hengsdijk
huib.hengsdijk@wur.nl

[1] Wageningen University & Research, P.O. Box 16, 6700 AA Wageningen, The Netherlands

Published online: 15 September 2017

Recognizing that thorough statistical analyses were hampered by lack of reliable data, the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) of the World Bank started to design and implement representative panel household surveys with a focus on agriculture. The systematically collected datasets contain information related to post-harvest management and self-reported post-harvest losses by farmers of different crops in addition to more general socio-economic information of farm households and geo-referenced information, for example on climate and distance of households to markets. In addition to these data, population weights are included and data are weighted to represent the national-level population of rural and small town households. Although eight African countries are involved in this project, in this study we focus on panel household data collected in Ethiopia. This multi-topic dataset allows analysis and modelling factors, which are associated with post-harvest losses in the literature such as the method of storage, ambient humidity and temperature, market access and household education (Tefera 2012; Stathers et al. 2013; Edoh Ognakossan et al. 2016).

The overall objective of this study is to disentangle factors that induce or relate to post-harvest losses of cereals in Ethiopia. A better understanding of the causes of post-harvest losses across geographical areas with different agro-ecological and socio-economic characteristics could enable more efficient targeting of interventions aimed at post-harvest loss reduction. This overall objective can be sub-divided into the following sub-goals:

1) To gain insight into the post-harvest storage management of cereals.
2) To gain insight into the scale, causes and reported percentages of post-harvest losses in cereal crops.
3) To identify and quantify major agro-ecological (e.g. altitude, rainfall, storage methods) and socio-economic (e.g. wealth of household, distance to market) variables that are related to post-harvest losses of cereals.

Kaminski and Christiaensen (2014) estimated post-harvest losses of maize in East Africa using LSMS-ISA data for Malawi (2010/2011), Tanzania (2008/2009 and 2010/2011) and Uganda (2009/2010). They concluded that on-farm self-reported post-harvest weight losses varied between 1.4 and 5.9% of the national maize harvest in these countries, while losses were concentrated among less than one fifth of the surveyed households. Kaminski and Christiaensen (2014) also analysed potential drivers of post-harvest losses, but the vast majority of the variation in post-harvest losses remained unexplained using classical parametric methods. In our study we started with a larger LSMS-ISA data set based on four East-African countries totalling seven years of data. Table 1 shows the number of cereal records (maize, barley, millet, rice, sorghum, teff, wheat) and the percentage of cereal records with post-harvest loss estimates (0%, >0%, and % missing data) in the available data set for Ethiopia, Malawi, Uganda and Tanzania. Post-harvest losses are recorded as such (in percentages) or calculated using the reported amount of crop losses (in local and/or SI units) divided by the total amount of harvested crop (in local and/or SI units) and expressed as percentage. In the remainder, we used post-harvest losses as a general term for both types of estimates. Except for the dataset of Ethiopia (2011/2012; 2013/2014) and Uganda (2011/2012) in roughly 90% of the cereal records the farmers' self-reported estimates of post-harvest losses were missing. The reason for the large number of missing data is unknown but made these survey data unsuitable for further analysis and modelling. The Ethiopia (2013/2014) and Uganda data were the most complete but indicated the prevalence of post-harvest losses in only 2% of the cereal records, which is also not helpful for modelling post-harvest losses. Therefore, in this paper we focused exclusively on the dataset of Ethiopia (2011/2012) as it contained the most complete data with quantitative post-harvest loss estimates and variables related to post-harvest management and losses. Also, different from Kaminski and Christiaensen (2014), we used Random Forests (RF) to model post-harvest losses of cereal crops.

Table 1 Number of available cereal records in various investigated country databases of the Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) and the percentage of cereal records with post-harvest loss values equal to 0%, > 0% and missing values

	Total # cereal records	Percentage of survey data records with estimates of post-harvest loss percentages		
		0%	> 0%	Missing value
Ethiopia (2011/2012)	5631	47	10	44
Ethiopia (2013/2014)	6290	96	2	2
Malawi (2010/2011)	2766	< 1	9	91
Tanzania (2008/2009)	2139		12	88
Tanzania (2010/2011)	2456		6	94
Tanzania (2012/2013)	3109		6	94
Uganda (2010/2011)	2968	82	2	16

RF is a non-parametric statistical ensemble learning method suited for the analysis of large data sets. Recently, RF has become popular in biological sciences because of its high-prediction accuracy and provision of information on the importance of variables for classification and regression (Breiman 2001; Touw et al. 2013).

In the remainder of this article we first describe the used data of Ethiopia (2011/2012), processing of data and the methods to model post-harvest losses. Section 3 describes 1) the post-harvest management of cereals, post-harvest losses and their causes in Ethiopia; 2) results of the post-harvest loss modelling for Ethiopia. In section 4 we reflect on the results of the various analyses, the RF method used to model post-harvest losses and the LSMS-ISA data used.

2 Data and methodology

2.1 Data: living standards measurement study-integrated surveys on agriculture (LSMS-ISA)

We used survey data of the LSMS-ISA project, which supports and collaborates with the national statistics offices of eight SSA countries (Burkina Faso, Ethiopia, Malawi, Mali, Niger, Nigeria, Tanzania, and Uganda) to design and implement systems of multi-topic, nationally representative panel household surveys with a focus on agriculture. The general setup of the surveys is the same across countries and typically consists of questionnaires related to the household, agriculture, livestock and community. The generic survey methodology carried out in different countries potentially allows cross-country and time series analyses of the data.

Household and post-harvest information in Ethiopia for 2011/2012 were collected over a period of 5 to 9 months, during which households were visited three times. The first round was in September–October 2011, the second round in November–December 2011 and the third round was in January–March 2012. The collected data were from the production year 2011. The LSMS-ISA Ethiopia data (2011/2012) covered all regional states except the capital, Addis Ababa, and formed a subset of the national agricultural sample survey. The LSMS-ISA survey was implemented in 290 rural and 43 small-town enumeration areas and includes all rural and small towns of Ethiopia except three zones of the Afar and six zones of the Somalia regions. The sample design provides representative estimates at the national level (excluding the nine zones in Afar and Somalia regions) for all rural households and for the combination of rural-area and small-town households. The regions of Ethiopia served as the strata of the two-stage sample design. Quotas were set for the number of enumerator areas in each region to ensure a minimum number of enumerator areas from each region. In the rural enumerator areas, a total of 12 households were sampled per enumerator area; 10 agricultural households were randomly selected from the agricultural sample survey, while the other two households were randomly selected. Population weights were included and applied to raise the sample households to national values for rural areas and small towns.

In addition to the standard survey questions relating to socio-economic variables of households, the survey also comprised information on the post-harvest characteristics of crop production, including self-reported quantitative estimates of post-harvest losses, self-reported causes of these losses and information on the post-harvest storage method and methods used to protect the cereals during the storage period. This information was used to gain insight into the post-harvest storage management and the scale, causes and self-reported percentages of post-harvest loss in major cereal crops; see Table 2 for an overview of the LSMS-ISA questions used for this purpose. The farmers' self-reported post-harvest loss estimates in the LSMS-ISA are considered to represent upper bounds when compared to storage and handling losses reported in the literature as they may include losses due to handling, drying, storage and marketing (Kaminski and Christiaensen 2014).

The households were further geo-referenced, using publicly accessible spatial databases. Information was provided on, for example, the distance to markets, annual rainfall and elevation. The survey data and detailed information on the sampling procedures, questionnaires, implementation of survey procedures and the spatial databases used can be found at the LSMS-ISA websites, accessible through www.worldbank.org.

2.2 Data processing

The LSMS-ISA project uses a number of topical questionnaires related to household, agriculture, livestock and community. We only used information collected through the household and agricultural questionnaires. The data were stored per topic in several files in which the households were identified by unique identifiers. A survey specific C# script was written, using the unique household identifier, to extract and collect all relevant information for data analysis on a per household basis. The newly prepared data file combined the information on post-harvest losses with the values of a number of variables. Post-harvest losses were recorded as such (percentages) or estimated using the amount of crop losses in IS units and the amount of harvested crop in IS units. For the Ethiopia data, 22 predictor variables were selected (Table 2), which can be categorized into demographic characteristics of the household head, such as age, gender and level of education; post-harvest management characteristics such as, cereal crop type, storage method and protection method; geo-referenced statistics, such as distance of the household dwelling to the nearest main road and nearest market, climate (average annual

Table 2 Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) data variables of Ethiopia (2011/2012) used 1) To characterize post-harvest management and post-harvest losses, 2) As potential predictor variables to model post-harvest losses, 3) In both type of analyses

Types of analyses:	LSMS-ISA survey variable:	LSMS-ISA survey questions and geo-variables:	Variable name in text	Unit / value(s) / name(s)
3	ph_s11q15_a	How much of the harvested [CROP] was lost to rotting, insects, rodents, theft, etc. in the post-harvest period?		kilo
3	ph_s11q15_b	How much of the harvested [CROP] was lost to rotting, insects, rodents, theft, etc. in the post-harvest period?		gram
3	ph_s11q15_c	How much of the harvested [CROP] was lost to rotting, insects, rodents, theft, etc. in the post-harvest period		percentage
3	ph_s9q12_a	How much [CROP] did you harvest from this field during the last completed agricultural season?		kilo
3	ph_s9q12_b	How much [CROP] did you harvest from this field during the last completed agricultural season?		gram
1	ph_s11q16_a	What was the reason for loss?		1) Rotting, 2) Insects, 3) Rodents/pests, 4) Flood, 5) Theft, 6) Other (specify)
3	Crop_code	Crop type	Crop	Barley, maize, millet, oats, rice, other cereals, sorghum, teff, wheat
3	ph_s11q18	What is your main method of storage for this crop?	Storage method	1) Unprotected pile, 2) Heaped in house, 3) Bags in house, 4) Metallic Silo, 5) Other (Specify)
3	ph_s11q20_a	What did you do to protect the stored [CROP]?	Protection1	1) Spraying, 2) Smoking, 3) Hired Guard, 4) Did Nothing, 5) Elevation, 6) Other (Specify)
2	ph_s11q21_a	In general, when you store [CROP], what is usually the main purpose for storing it?	Purpose	1) for household consumption, 2) to sell at a higher price, 3) seed for planting, 4) render payments in-kind, 5) wait for the arrival of buyer, 6) does not usually store, 7) emergency, 8) Other (specify)
2	ph_s11q22_a	What proportion of your crop have you used for household consumption?	Consumption	%
3	hh_s1q02	What is the sex of [NAME]? (of the household head)	Gender	1) male, 2) female
3	hh_s1q04_a	How old is [NAME]? (the household head)	Age	number
3	hh_s2q02	Can you read and write in any language?	Ability to read/write	1) yes, 2) no
2	hh_s2q03	Have you ever attended school?	Attended school	1) yes, 2) no
3	dist_road	Household distance to nearest major road	Distance to main road	km
3	dist_market	Household distance to nearest market	Distance to nearest market	km
3	af_bio_1	Average annual temperature calculated from monthly climatology	Annual mean temperature	°C * 10
2	af_bio_12	Total annual precipitation (from monthly climatology)	Annual precipitation	mm
3	anntot_avg	Average 12-month total rainfall(mm) for January–December (2001–2011)	Average rainfall	mm
2	h2011_tot	12 month total rainfall in January–December, starting January 2011	Average rainfall 2	mm
2	h2011_wetQ	Total rainfall in wettest quarter within 12 month periods, starting January 2011	Total rainfall	mm
2	twi	Topograhic wetness index	Wetness index	Number
2	srtm	Elevation	elevation	m
2	LAT_DD_MOD	Latitude (WGS84) of Enumeration area	latitude	Number
2	LON_DD_MOD	Longitude (WGS84) of Enumeration area	longitude	Number
2	hh_s9q19		Light household	

Table 2 (continued)

Types of analyses:	LSMS-ISA survey variable:	LSMS-ISA survey questions and geo-variables:	Variable name in text	Unit / value(s) / name(s)
		What is the main source of light for the household?		1) electricity meter – private, 2) electricity meter – shared, 3) Electricity from generator, 4) solar energy, 5) bio-gas, 6) electrical battery, 7) lantern, 8) light form dry cell with switch, 9) kerosene light lamp imported, 10) local kerosene lamp (Kuraz), 11) candle/wax, 12) fire wood, 13) Other (specify)
2	hh_s9q21	What is the main source of cooking fuel?	Fuel household	1) collecting fire wood, 2) purchase fire wood, 3) charcoal, 4) crop residue/leaves, 5) dung/manure, 6) sawdust, 7) kerosene, 8) butane – gas, 9) electricity10) solar energy, 11) bio-gas, 12) none, 13) Other (specify)
3	pw	household sample weight		–

rainfall, total rainfall in 2011, annual mean temperature) and elevation; wealth status of the household approximated by the type of energy source for cooking (e.g. electricity vs. fire wood) and by the main source of light for the household (e.g. electricity meter vs. kerosene light lamp). Prior to the analysis, population weights for households were normalized to uniform weights. Although the original questionnaires contained many more variables, only those variables for which an association with post-harvest losses could be expected were selected for analysis with RF.

To assess participation bias, a nonresponse analysis was performed for the demographic characteristics, geo-referenced statistics and wealth status. For these variables, mean values were compared between the base cohort (5631 records) and the post-harvest loss cohort containing non-missing post-harvest losses only (3179 records).

2.3 Methods

To model post-harvest losses of cereals in Ethiopia we used Random Forests (RF), which is based on regression trees. A regression tree is a predictive modelling approach where many variables are mapped on a tree-like structure to predict a target value (Breiman et al. 1984). The outcome of the model can be graphically displayed as a binary tree showing how the dependent variable is affected by the predictor variables. The tree consists of nodes and branches and, depending on the value of a predictor variable at each node, one of two sub-branches is followed, finally ending in a leaf, which represents the target variable. The tree is grown using a training set and, at the same time, an independent set of data is mapped on this tree to evaluate the model.

Random Forests differs from regression trees in that not a single tree is grown but a large number of uncorrelated trees. This is the so-called forest. Each tree in the forest is grown using a bootstrap sample of the data. A vector of non-negative weights containing uniform probabilities is used to select cases as candidates for the bootstrap. After a large number of trees, the predicted value becomes available as the combined results across all trees using the cases that are not in the bootstrapped set. A difference with standard regression trees where a node is split using the best split among all predictor variables, is that in RF, a node is split using the best split among a random subset of input variables. Generating a forest of trees using bootstrapping in combination with random selection of predictor variables has several advantages compared to standard regression trees. In each bootstrap iteration a tree is grown using the training set and, at the same time, predicted values become available using the independent data. There is no need to prune, trees are grown very deep but variance is reduced by averaging many trees. One of the drawbacks of a random forest is that some interpretability is lost but, in general, the performance of the final model is boosted (Breiman 2001).

Random Forests can effectively handle large datasets that contain many variables with complex relationships. Though initially developed to maximize the predictive performance of the model, RF has a number of methods available for exploratory data analysis and interpretation of complex nonlinear relationships between explanatory and outcome variables. Graphical methods, like partial dependence plots, extract this information and visualize the relation between predictor variable and outcome. Variable importance scores can be calculated that indicate the relevance of a predictor variable for the outcome of the model. Values close to zero indicate that the variable is not important, where high values indicate that the predictive power of the forest is improved by including them.

When used for prediction only, there is no need to remove non-informative variables. In this study, where the aim is to improve understanding of the determinants of post-harvest

losses, the number of variables in the model was reduced based on importance scores to improve the interpretation and understanding of factors contributing to post-harvest losses. Partial dependence plots have been used to visualize these relations. The partial effect of a variable was constructed for a range of evenly spaced values of the variable of interest, while keeping the values of the other variables unchanged. By taking the average prediction of the RF over all other covariates in such a point, the conditional effect of the variable of interest was calculated. Partial dependence co-plots are useful for investigating the combined effect of two variables on the response or to visualize pairwise interaction effects among variables. For a categorical variable, the partial effect of a continuous variable was calculated, conditional on the group membership. For continuous data, conditional membership has been accomplished by stratifying the conditioning variable into subgroups. For the relevant variables of the LSMS-ISA data, two groups of equal size were created each one representing a group with respectively values below and above the median value. Then, the partial effect of a variable was calculated conditional on the membership of the high and low group of the grouping variable.

Random Forests is available in a number of R packages. We used the **randomForest** package (Liaw 2015; Liaw and Wiener 2002, available at http://cran.r-project.org/package=randomForest) and the **randomForestSRC** package, version 2.4.1 (Ishwaran and Kogalur 2014; available at http://cran.r-project.org/package=randomForestSRC). The last package introduces the **ggRandomForest** package (Ehrlinger 2015), which implements tools for extracting intermediate data objects from the **randomForestSRC** package and uses the **ggplot2** graphics package (Wickham 2009) to visualize RF models.

3 Results

3.1 General household information

Table 3 describes the major characteristics of the surveyed households in Ethiopia, which totalled 2472 unique households with cereals. Since most households managed several plots, the analysis comprises information of 5631 cereal plot records. From these records 3179 plots self-reported post-harvest losses of 0% or higher were reported. In Table 3, the means for the major variables are shown, both for the base cohort (n = 5631) and for the post-harvest loss cohort with reported values only (n = 3179; 56%). As shown, mean values for age, female headed households, annual rainfall, distance to the main road and nearest market were generally similar for both cohorts, demonstrating that the post-harvest cohort was an unbiased sample from the base cohort for cereals.

Table 3 Characteristics of households in Ethiopia (2011/2012) from the Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) used to characterize post-harvest management and post-harvest losses

	base cohort		post-harvest cohort	
	number of records	% of records	number of records	% of records
Number of households	2472		1402	
Total cereal records:	5631		3179	
Barley	633	12%	356	12%
Maize	1740	29%	1000	30%
Millet	333	6%	188	6%
Oats	11	< 1%	10	< 1%
Other cereals	2	< 1%	0	< 1%
Rice	18	< 1	9	< 1%
Sorghum	1206	18%	711	21%
Teff	1054	21%	572	19%
Wheat	634	13%	333	11%
Average age of household head		44 years		44 years
Female headed households		14%		14%
Illiteracy of household heads		59%		58%
Average distance to nearest market		58 km		59 km
Average distance to main road		15 km		15 km
Average annual rainfall		905 mm		895 mm
Average annual temperature		18.7 °C		18.4 °C

The base cohort contains all cases including cases with missing values for post-harvest losses. The post-harvest cohort contains only cases with non-missing values for post-harvest losses. All statistics are weighted using household sample weights

Maize is the major cereal crop reported on 29% of all cereal records, while sorghum (18%) and teff (21%) are also major cereals. Wheat and barley together make up 25% of the cereal records, while millet and especially the number of records with oats, rice and other cereals are negligible.

The average age of the household heads was 44 years, and 14% of the households were headed by females. The majority of the households in Ethiopia were illiterate (59%). Households live far from the nearest markets, on average 58 km. The distance to the main road is on average 15 km. The annual rainfall is about 905 mm and temperature is 18.7 °C.

3.2 Post-harvest storage management characteristics

Table 4 shows the different storage methods used in Ethiopia for the major cereals: maize, sorghum, teff, wheat and barley. Differences in storage method among the different types of cereals were relatively small. The most important method of storing cereals was a bag in the house; about 46% of the cereal records are stored this way. Modern storage methods, such as metal silos were hardly used. 'Other' storage methods, 39% of all responses, refer probably for a large part to the widely used traditional *Gotera*. This is an elevated storage platform often made from locally available material on which grains are stored close to dwellings. In the LSMS-ISA survey of Ethiopia 2013/2014 the option 'traditional storage' was added as a possible response and accounted for about 26% (unweighted) of all cereal storage methods (data not shown).

Table 5 shows the different protection methods applied by farmers to reduce storage losses. The most frequently used protection measure for cereals in Ethiopia was elevation (for 43% of all cereals), which relates to the traditional *Gotera* storage platform (see previous paragraph). Elevation as a protection method was especially used for teff, wheat and barley. Traditional and modern pesticides tended to be more frequently used in maize and sorghum with 32 and 30%, respectively of all protection methods used in these crops. At least 30% of the respondents did not use any protection method, while another 26% did not provide a response on the method used for protection.

3.3 Post-harvest losses and causes

Table 6 shows the average percentage of self-reported post-harvest loss estimates per cereal type and cause. These estimates are restricted to records with losses higher than 0% ($n = 529$). As shown in Table 1, only 10% of the cereal records contained losses higher than 0%, therefore, some of the average losses shown in Table 6 are based on a few estimates only. Given this limitation, the average self-reported post-harvest loss over all cereals was 24% with a somewhat higher loss for wheat, 27% and a lower loss for teff, 21%. The average post-harvest loss estimate due to 'other' factors was highest with 35%. The average post-harvest loss estimate due to insects and rotting was 27% for both, while for theft the lowest average loss was 5%.

Table 7 shows the frequency of self-reported causes of post-harvest loss expressed as the percentage of the total number of reported causes of loss per cereal ($n = 529$). Rodents and other pests were most frequently reported as causes of post-harvest losses, on average 46%. The highest percentage was found for maize, 52%, the lowest for teff, 32%. Only for teff 'other' causes of post-harvest loss (36%) were more frequently reported causes.

3.4 Modelling post-harvest losses

The Ethiopian records with self-reported post-harvest losses were analysed (n = 3179), records with missing values and post-harvest losses >100% were excluded. Missing values in the data matrix were imputed using the proximity matrix from the forest. The number of trees was 500 and the initial number of variables tried at each split was 7. After imputation, a full random forest was grown (ntree = 1000). The percentage of explained variance was 31%.

Figure 1 shows the importance scores, normalized by the standard deviation, of the predictor variables in the model for

Table 4 Post-harvest storage methods in Ethiopia (2011/2012) expressed as percentage of the total number of storage methods used for each cereal type without taking into account the records with missing information on storage methods

	Bags in House	Heaped in House	Metallic Silo	Unprotected Pile	Other	Missing data	Number of records base cohort
Maize	42	7	< 1	7	43	30	1740
Sorghum	38	9	1	6	46	14	1206
Teff	47	7	< 1	9	37	15	1054
Barley	58	8	0	6	29	21	633
Wheat	61	6	0	4	29	15	634
Average	46	7	< 1	7	39	20	

The number of records (far right column) represents the base cohort for major cereals ($n = 5267$) containing all records including records with missing values for post-harvest losses. All statistics are weighted using household sample weights

Table 5 Protection methods used by farmers in Ethiopia (2011/2012) during the storage period

	Did Nothing	Elevation	Spraying[a]	Smoking	Hired Guard	Other	Missing data	number of records base cohort
Maize	23	39	32	4	0	3	43	1740
Sorghum	29	38	30	1	0	2	16	1206
Teff	37	52	6	1	0	4	18	1054
Barley	36	49	10	2	0	3	25	633
Wheat	30	44	19	2	0	5	16	634
Average	30	43	22	2	0	3	26	

Protection methods expressed as percentage of the total number of methods used for each cereal type without taking into account the records with missing information on protection methods. The number of records (far right column) represents the base cohort for major cereals (n = 5267) containing all records including records with missing values for post-harvest losses. All statistics are weighted using household sample weights

[a] 'Spraying' is used in the English questionnaire (Table 2). The Amharic in the questionnaire was "spraying medicine". Medicine is a general term for a traditional or modern pesticide, while spraying could be understood as "adding", "mixing with" or "dusting"

Ethiopia. Gender, age and variables related to education have low scores and are not useful for explaining post-harvest loss. Scores for crop type and methods of protection during storage were slightly higher but their contribution to the predictive power of the model was also negligible. Geo-referenced variables such as latitude, distance of the household dwelling to the main road and nearest market and different variables used to characterise rainfall are important and informative in describing post-harvest losses.

In the full model, post-harvest losses were modelled using 22 predictor variables. To end up with a set of variables that facilitates understanding of the causes of post-harvest loss, we removed variables from the model, one at a time, starting with the variable with the lowest importance score. Then, the random forest was grown again and the next variable with the lowest importance score was removed. This process was repeated until the percentage of explained variance dropped substantially. The reduced model contained four variables and the percentage of explained variance finally dropped from 31 to 27%. Distance from the household to the main road and nearest market, average rainfall and latitude are important determinants of post-harvest losses for the Ethiopian data. Because latitude did not add to the interpretation of the model and because some level of confounding with distance to nearest market and main road and average annual rainfall could be present, we dropped this variable. The percentage of explained variance dropped to 26%, meaning that the predictive value of the model was not significantly influenced by leaving out latitude. Figure 2 shows the importance scores of the reduced model. In the RF analysis all cereals were pooled, despite the fact that crops may behave differently during storage and that pests attacking the harvest may differ among crops. Recognizing this, the variable crop type was used as a predictor variable in the model. The importance score for the variable crop type indicates that incorporating this variable in the model does not contribute to the explanation of the variability in post-harvest losses and that this variable could equally well be dropped (Fig. 1).

To interpret the random forest, partial dependence plots have been made to visualize the individual effect of the three continuous predictor variables on post-harvest losses, i.e. distance of the household dwelling to main roads and nearest

Table 6 Average percentage of self-reported post-harvest loss (%) in Ethiopia (2011/2012) per cereal type and cause (n = 529

	Insects	Rodents/Pests	Rotting	Flood	Other	Theft	Average	number of records
Maize	37	16	26	25	40	1	24	208
Sorghum	13	20	36	21	29	22	23	126
Teff	7	14	18	30	33	0	21	76
Barley	23	12	24	31	39	0	23	78
Wheat	55	12	30	1	34	1	27	41
Average	27	15	27	22	35	5	24	

For example, farmers with sorghum reported an average post-harvest loss of 13% due to insects, while the average loss reported by all farmers with sorghum was 23%. The set of post-harvest loss estimates is restricted to losses higher than 0%. All statistics are weighted using household sample weights

Table 7 The frequency of self-reported causes of post-harvest loss (> 0%) in Ethiopia (2011/2012) expressed as percentage of the total number of reported causes per crop type (n = 529)

	Insects	Rodents/Pests	Rotting	Flood	Others	Theft	number of records
Maize	16	52	12	2	18	1	208
Sorghum	10	47	15	2	20	6	126
Teff	4	32	26	2	36	0	76
Barley	9	43	17	5	27	0	78
Wheat	20	40	12	3	24	2	41
Average	12	46	15	3	23	2	

The set of post-harvest loss estimates is restricted to losses higher than 0%. For example, for maize, 16% of the 208 reported records the cause was insects. All statistics are weighted using household sample weights

markets and average annual rainfall (Fig. 3). For the distance of the household to both the nearest market and the main road, the trend is that post-harvest losses are higher when the distance increases. The linear trend for average annual rainfall shows a negative effect on post-harvest losses. However, the individual data points on both extremes of the plot suggest that losses at low (< 600 mm) and high (>1200 mm) annual rainfall may be higher.

Because interactions are likely to occur, the combined effect of the predictor variables is of real interest. To get a first idea about the size and direction of the effect, the values of distance to the main road and nearest market were divided into two groups of equal size using the median as splitting value yielding groups with respectively low and high values. For each group, the partial effect on post-harvest loss of the other variable was estimated and plotted and the trend was indicated by a linear function through the estimated values for post-harvest loss. Figure 4 shows the partial effect of distance of households to the nearest market conditional on the low (< 23 km) and high (> 23 km) group for distance of households to the main road. As also shown in Fig. 3 self-reported estimates of post-harvest losses increase for households located further away from the nearest market. There is hardly any effect on the self-reported post-harvest losses for households living close to the main road (continuous line in Fig. 4) or households living further away (dotted line in Fig. 4).

In Fig. 5 the partial effect of distance of households to the main road conditional on the low (< 63 km) and high (> 63 km) group for distance of households to the nearest market is shown. Estimated post-harvest losses increase to the same extent both for households living further away from the main road (> 63 km, dotted line in Fig. 5) and households with a relatively nearby market (< 63 km, continous line in Fig. 5).

In Fig. 6 the partial effect of average annual rainfall conditional on the low (< 23 km) and high (> 23 km) group for distance of households to the main road is shown. Both types of households living further away from a main road and living nearby a main road reported higher post-harvest losses under low rainfall conditions. There is no interaction effect, i.e. the effect of rainfall on self-reported post-harvest losses is not influenced by the distance of households to the main road.

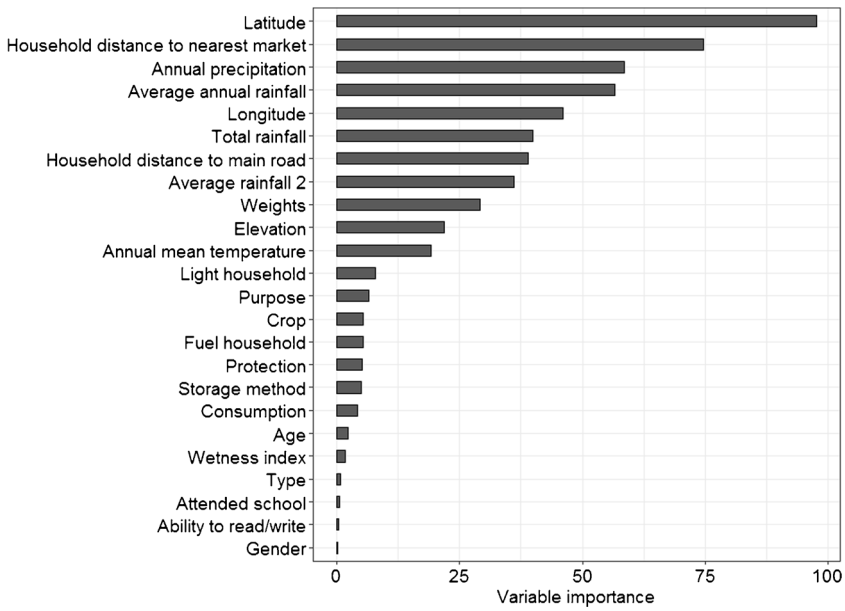

Fig. 1 The importance scores of variables in explaining the self-reported post-harvest loss of cereals in Ethiopia (2011/2012)

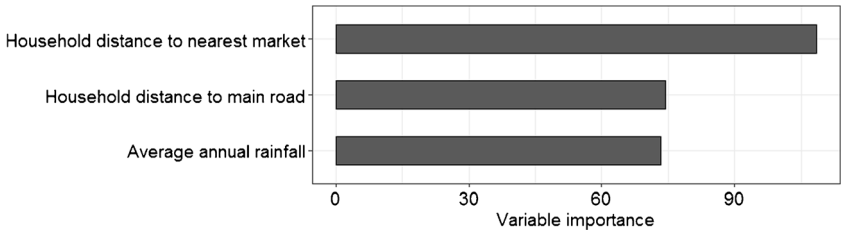

Fig. 2 The importance of the main variables in explaining the self-reported post-harvest loss of cereals in Ethiopia (2011/2012)

Figure 7 shows the partial effect of average annual rainfall conditional on the low (< 63 km) and high (> 63 km) group for distance of households to the nearest market. Both types of households reported higher post-harvest losses under low rainfall conditions independent of the distance to the nearest market.

4 Discussion and conclusions

We started with analysing LSMS-ISA national survey data of more than 15,000 households and more than 25,000 cereal records from four countries (Ethiopia, Malawi, Tanzania and Uganda) and covering seven years to gain better insight into

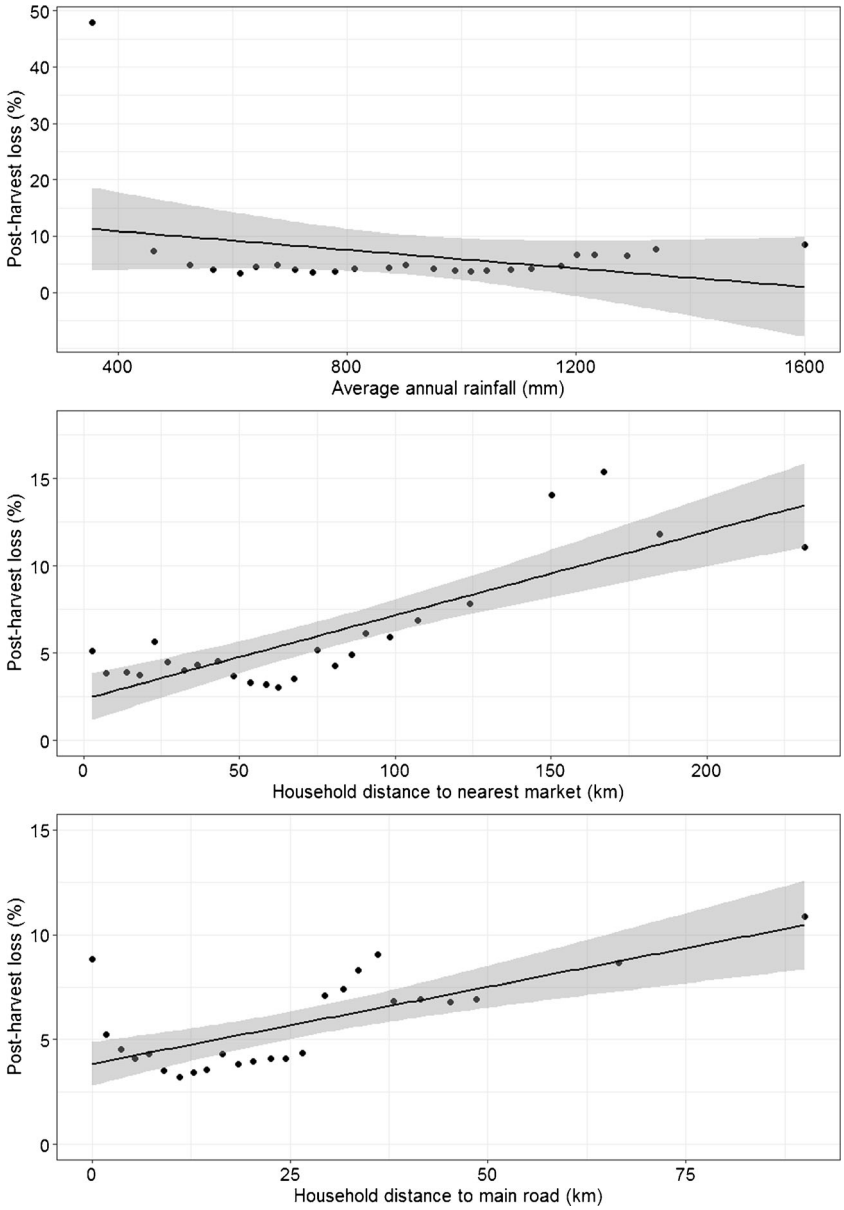

Fig. 3 Partial dependence plots for the major predictor variables for post-harvest loss (%) in cereals in Ethiopia (2011/2012): distance to nearest markets, distance to main road and average annual rainfall. The shaded areas around the lines indicate the 95% confidence interval

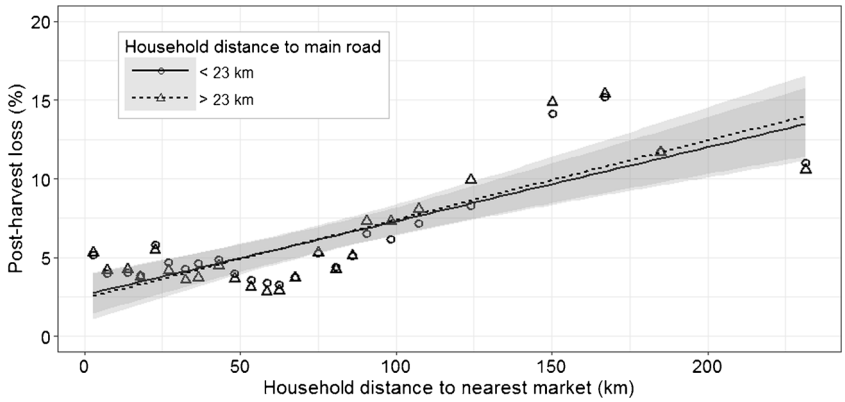

Fig. 4 The partial effect on post-harvest losses (%) in cereals of the distance of households to the nearest market (km) conditional on the low (< 23 km) and high (> 23 km) group for distance of households to the main road (km) in Ethiopia (2011/2012). The shaded areas around the lines indicate the 95% confidence interval

post-harvest storage management and post-harvest losses of cereals in SSA. However, in the four datasets of Tanzania and Malawi about 90% of the responses on the farmers' self-reported post-harvest losses in cereals were missing (Table 1). Such large amounts of missing data do not allow meaningful modelling and analysis of post-harvest losses, and highlight extreme problems in the data collection and quality management of these large survey data sets. The datasets of Ethiopia (2013/2014) and Uganda were most complete but indicated the prevalence of post-harvest losses in only 2% of the cereal records, which is also not helpful to for modelling and better understanding post-harvest losses. Therefore, in this paper we focussed exclusively on the dataset of Ethiopia (2011/2012). Yet, 44% of the self-reported post-harvest loss estimates were missing in this dataset, but as shown in Table 3 the difference between the sample with post-harvest data and the base cohort (including missing values) was negligible. Therefore, our findings and conclusions are representative for the entire data set of Ethiopia (2011/2012).

The LSMS-ISA data could potentially have a number of advantages over the use of other post-harvest loss data sources such as case study data and expert estimates of losses (Kaminski and Christiaensen 2014): (i) sample bias is avoided because the survey data provide nationally-representative samples of agricultural households and the post-harvest losses these households report; (ii) harmonization in the survey methodology facilitates comparison of the outcomes across years and countries. However, the main reason for using the LSMS-ISA data in our study is that the multi-topic and geo-referenced survey approach helps to improve our understanding of those agro-ecological factors (e.g. altitude, rainfall, storage methods) and socio-economic conditions of households (e.g. wealth of household, distance to market) that favour post-harvest losses. This helps to better target interventions aimed at reducing post-harvest losses. One disadvantage of the LSMS-ISA data is that the post-harvest-loss estimates are based on subjective reported information from farmers, which may be less accurate than measured loss data. However, practical, methodological as well as conceptual challenges to measure accurately post-harvest losses at farm level are great (Parfitt et al. 2010; Hodges 2013; Affognon et al. 2015). Another disadvantage of the LSMS-ISA data is that the current post-harvest loss estimate is an aggregated loss of all possible losses that may occur during the entire post-harvest chain. More detailed information on where losses in the post-harvest chain occur would be useful to better target interventions aimed at reducing such losses. Understanding of post-harvest management and losses can be increased considerably through adding survey questions on the post-harvest losses incurred during different stages of the post-harvest chain, such as harvesting, drying, winnowing and storage. The currently available post-harvest loss estimates from LSMS-ISA data

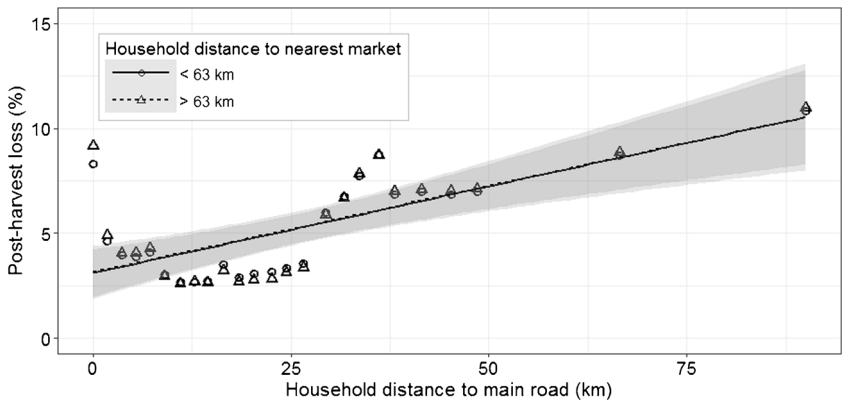

Fig. 5 The partial effect on post-harvest loss (%) in cereals of the distance of households to the main road (km) conditional on the low (63 km) and high (> 63 km) group for distance of households to the nearest market (km) in Ethiopia (2011/2012). The shaded areas around the lines indicate the 95% confidence interval

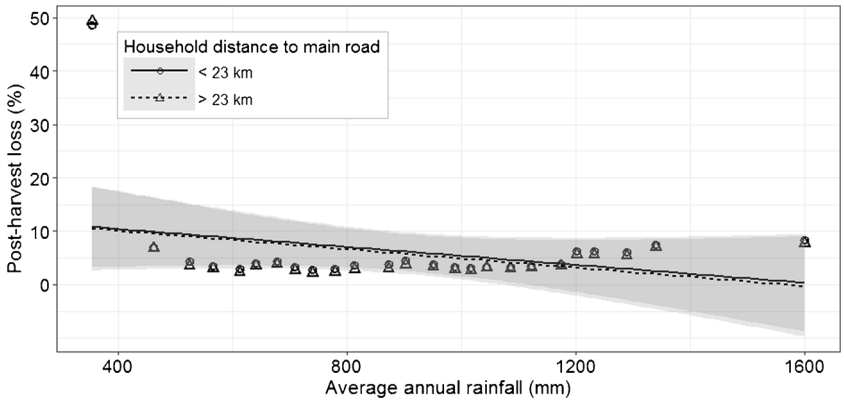

Fig. 6 The partial effect on post-harvest losses (%) in cereals of average annual rainfall (mm) conditional on the low (< 23 km) and high (> 23 km) group for distance of households to the main road (km) in Ethiopia (2011/2012). The shaded areas around the lines indicate the 95% confidence interval

discloses losses that farmers regard as important and therefore provide an appropriate yardstick for assessing losses that are imperative for reduction through targeted interventions.

As indicated at the beginning of this section, the major reason for not using more LSMS-ISA data were the large amount of missing data with respect to self-reported post-harvest loss estimates, especially in the datasets of Tanzania and Uganda. The reasons for the large amount of missing data are unknown and not limited to only post-harvest management and loss variables. Other agricultural and household data that were not used in our study also showed large data gaps, which raises doubts about the survey implementation and data quality control. The Ethiopian data set (2011/2012) appeared to be the most complete and allowed an insight into current post-harvest management of cereals to be gained through Random Forests analysis. This method disentangled those factors that induce or relate to post-harvest losses of cereals in Ethiopia. More than 50% of the cereals in Ethiopia are stored inside the house in bags or piles. Also traditional elevated storage platforms (*Goteras*) close to the dwellings are still frequently used. Elevation was the most used protection method, while traditional and modern pesticides tended to be more frequently used in maize and sorghum than in teff, barley or wheat. The farmers reporting post-harvest loss estimates faced an average weight grain loss of 24%, within a relatively small range of a lowest loss of 21% for teff and a highest loss of 27% for wheat (Table 6). Storage losses of teff are known to be lower than other cereals because of the small grain size, which makes teff more resistant to insect attacks than other cereals (WB/NRI/FAO 2011). Most frequently reported causes of post-harvest losses were rodents and other pests, except for teff for which 'other causes' were more important (Table 7). The high post-harvest loss estimates in cereals due to 'other causes' (35%; Table 6) and the high frequency of reported 'other causes' of post-harvest losses (23%; Table 7) calls for more in-depth research into these causes.

In the modelling of post-harvest losses of cereal crops using Random Forests we pooled the data for cereal crops. Although the pooling of all cereal crops in one analysis may be criticized, the outcome of the Random Forest shows that post-harvest losses did not depend significantly on the type of crop (Fig. 1). The crop variable was in the first set of 22 predictor variables, but the importance score was low. After dropping crop type as a predictor variable from the model, the percentage of explained variance did not drop significantly, indicating that the type of crop was not important in explaining post-harvest losses.

RF is well suited for the analysis of large, noisy data sets that exhibit highly irregular patterns, nonlinear effects and interacting variables. Moreover, RF can easily handle many predictor variables that may be correlated or not, have interactions or not and do not require any distributional

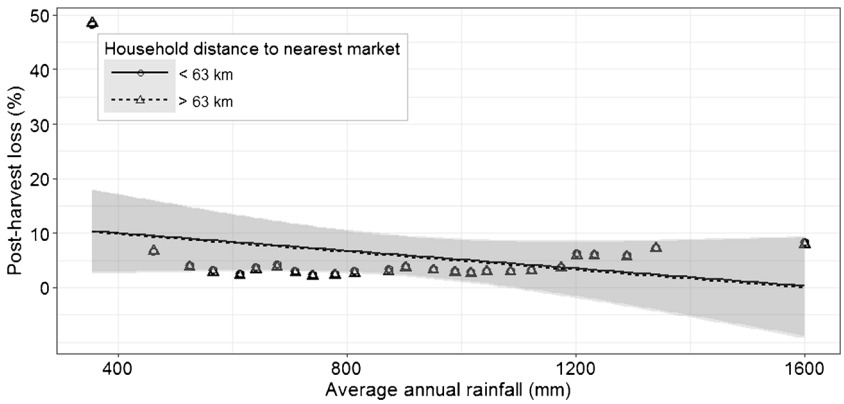

Fig. 7 The partial effect on post-harvest loss (%) of average annual rainfall (mm) conditional on the low (< 63 km) and high (> 63 km) group for distance of households to the nearest market (km) in Ethiopia (2011/2012). The shaded areas around the lines indicate 95% confidence interval

assumptions. Classical approaches to model such data, such as ANOVA, regression, or (generalized linear) mixed models are often hampered by the high number of potentially available predictor variables that interact in unexpected ways, making it hard to identify important determinants or combinations of determinants of post-harvest losses. Complex nonlinear relationships are often missed without explicit pre-specification thus complicating the disclosure of new features of the data and may easily lead to spurious conclusions (Kaminski and Christiaensen 2014; Krupnik et al. 2015). This is also confirmed by Flack and Chang (1987), who demonstrated that variable selection within a set of noisy predictor variables frequently resulted in selected subsets of noise variables. Therefore, RF was a flexible method to explore the Ethiopian data. The initial model for the Ethiopian dataset (2011/2012) contained 22 variables and explained 31% of the variance. Many variables were correlated compromising the interpretation of the model. For example, for describing precipitation, the LSMS-ISA data used four variables, all characterising different aspects of the amount of rainfall during a year. The RF algorithm turned out to be flexible enough in dropping variables that were confounded without losing too much of its performance. By reducing the model, its interpretability was enlarged at the expense of losing some predictive power. Therefore, the results as derived with RF, may be considered as the best attempt to obtain information embedded in the data. Because of the advantages mentioned above, exploratory data analysis with RF will probably outperform classical approaches such as regression.

The percentage variance of the final model explained was low for the Ethiopian 2011/2012 data. Nevertheless, this dataset can be considered as an illustration as to how RF can be used to analyse such large data sets and how methods such as partial dependence can be used to extract substantive insights from the forest. In the Ethiopia 2011/2012 data the distance of the household dwelling to the nearest market and main road, and the average annual rainfall were identified as major factors that affected post-harvest losses in cereals. In Ethiopia, households living further away from markets and main roads report the highest post-harvest losses, while lower rainfall (higher losses) had a minor effect compared to the remoteness of households. Therefore, infrastructure and access to markets is not only of major importance for stimulating agricultural productivity, growth and development but also for reducing post-harvest losses (Dorosh et al. 2012; Tefera 2012). Thus, the reduction of post-harvest losses in Ethiopia requires large public investments but these are complementary to investments required for achieving productivity growth and food security (Rosegrant et al. 2015).

The LSMS-ISA infrastructure is well placed within national statistical agencies to collect through surveys information on post-harvest management and self-reported post-harvest loss information across the range of crops grown in various SSA countries. This information is potentially suitable to model post-harvest losses identifying generic factors and conditions favouring losses. However, the available LSMS-ISA information on both storage management and self-reported post-harvest losses shows that implementation of the surveys differ greatly across countries. Overall, information on crop storage, protection methods used during the storage period and above all, on the self-reported post-harvest loss estimates is incomplete in various LSMS-ISA data sets. This hampers the identification and quantification of important variables and conditions associated with post-harvest losses in SSA, which can help to identify appropriate interventions to reduce post-harvest losses. More emphasis should be placed on improving the quality, relevance and use of data and checking data at the early stages of data collection. This paper is therefore a call for greater awareness raising on the importance of post-harvest management and losses at every level but also a call for better data collection for which the infrastructure is already in place.

Acknowledgements We thank the staff of the World Bank for answering questions related to the LSMS-ISA data. In addition, we thank four reviewers for their valuable comments on earlier versions of this paper, which helped us to better focus the paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Affognon, H., Mutungi, C., Sanginga, P., & Borgemeister, C. (2015). Unpacking postharvest losses in sub-Saharan Africa: a meta-analysis. *World Development, 66*, 49–68.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Breiman L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey: Wadsworth and Brooks.

Dorosh, P., Wang, H. G., You, L., & Schmidt, E. (2012). Road connectivity, population, and crop production in sub-Saharan Africa. *Agricultural Economics, 43*, 89–103.

Edoh Ognakossan, K., Affognon, H.D., Mutungi, C.M., Sila, D.N., Midingoyi, S-K.,G., Owino, W.O. (2016). On-farm maize storage systems and rodent postharvest losses in six maize growing agro-ecological zones of Kenya. Food Security, 8, 1169–1189.

Ehrlinger, J. (2015). ggRandomforests: random Forest for regression. http://cran.r-project.org/web/packages/randomForest. Accessed 2 May 2016.

FAO [Food and Agriculture Organization]. (2011). Global food losses and food waste. Study conducted for the International congress SAVE FOOD! At Interpack 2011, Düsseldorf.

Flack, V. F., & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: a simulation study. *The American Statistician, 41*, 84–86.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulimin, C. (2010). Food security: the challenge of feeding 9 billion people. *Science, 327*, 812–818.

Hertel, T. W. (2015). The challenges of sustainably feeding a growing planet. *Food Security, 7*, 185–198.

Hodges, R.J. (2013). How to assess postharvest cereal losses and their impact on grain supply: rapid weight loss estimation and the calculation of cumulative cereal losses with the support of APHLIS. Version 1.1. http://www.aphlis.net/downloads/APHLIS%20Losses%20Manual%2013%20Dec%2013%20revised.pdf. Accessed 15 November 2016.

Hodges, R. J., Buzby, J. C., & Bennett, B. (2011). Postharvest losses and waste in developed and less developed countries: opportunities to improve resource use. *Journal of Agricultural Science, 149*, 37–45.

Ishwaran H., Kogalur U.B. (2014). Random Forest for survival, regression and classification (RF-SRC), R package version 2.6. http://cran.r-project.org/package=randomForestSRC. Accessed 2 May 2016.

Kaminski, J., & Christiaensen, L. (2014). Post-harvest loss in sub-Saharan Africa – what 1 do farmers say? *Global Food Security, 3*, 149–158.

Krupnik, T. J., Ahmed, Z. U., Timisina, J., Yasmin, S., Hossain, F., Mamun, A. A., Mridha, A. I., & McDonald, A. J. (2015). Untangling crop management and environmental influences on wheat yield variability in Bangladesh: an application of non-parametric approaches. *Agricultural Systems, 139*, 166–179.

Liaw, A. (2015). R randomForest package documentation. Available at: http://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Accessed 7 December 2015.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2*, 18–22.

Lundqvist, J., De Fraiture, C., & Molden, D. (2008). Saving water: from field to fork – curbing losses and wastage in the food chain. Stockholm: SIWI policy brief, the Stockholm International Water Institute (SIWI).

Parfitt, J., Barthell, M., & Macnaughton, S. (2010). Food waste within food supply chains: Quantification and potential change to 2050. *Philosophical Transactions of the Royal Society. Biological Sciences, 365*, 3065–3081.

Rembold, F., Hodges, R., Bernard, M., Knipschild, H., & Leo, O. (2011). The African postharvest losses information system (APHLIS). EUR 24712 EN-Joint research centre-institute for environment and sustainability. Luxembourg: Publications Office of the European Union.

Reynolds, T. W., Waddington, S. R., Anderson, C. L., Chew, A., True, Z., & Cullen, A. (2015). Environmental impacts and constraints associated with the production of major food crops in sub-Saharan Africa and South Asia. *Food Security, 7*, 795–822.

Rosegrant, M. W., Magalhaes, E., Valmonte-Santos, R. A., & Mason-D'Croz, D. (2015). Returns to investment in reducing postharvest food losses and increasing agricultural productivity growth: Post-2015 Consensus. Food Security and Nutrition Assessment Paper. Lowell: Copenhagen Consensus Center.

Stathers, T., Lamboll, R., & Mvumi, B. M. (2013). Postharvest agriculture in changing climates: Its importance to African smallholder farmers. *Food Security, 5*, 361–392.

Tefera, T. (2012). Post-harvest losses in African maize in the face of increasing food shortage. *Food Security, 4*, 267–277.

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2013). Datamining in the life sciences with random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics, 14*, 315–326.

WB/NRI/FAO [World Bank/Natural Resources Institute/Food and Agriculture Organization] (2011). Missing food: the case of postharvest grain losses in sub-Saharan Africa. Report no. 60371-AFR. Washington, DC.

West, P. C., Gerber, J. S., Engstrom, P. M., Mueller, N. D., Brauman, K. A., Carlson, K. M., Cassidy, E. S., Johnston, M., MacDonald, G. K., Ray, D. K., & Siebert, S. (2014). Leverage points for improving food security and the environment. *Science, 345*, 325–327.

Wickham H. (2009). ggplot2: Elegant graphics for data analysis. Springer New York. ISBN 978-0-387-98140-6.

Huib Hengsdijk is an agronomist at Wageningen University and Research (WUR) focussing on research and development related to agriculture and natural resources management. He has worked in Central America, Southeast Asia and Africa. A common denominator of his work is the assessment and exploration of improving resource use efficiencies in agricultural systems. He has a background in agro-systems modelling at different scales from assessing crop yield variability at plot level, designing innovative and feasible cropping systems at farm level to exploring food production potentials at regional level. Recently, he also has been working on linking smallholders to value chains and capacity building and training of extension and farmers in good agricultural practices in both Ethiopia and Indonesia. Currently, his work focuses on the use of geo-data for agriculture through crop insurance for rice in Indonesia and the control of late blight in potatoes in Bangladesh.

Waldo de Boer is a statistican at Biometris, Wageningen University and Research. He is an experienced advisor on statistics applied to food safety, biological issues and statistical process control. Over the last 15 years, Waldo has been working in large European projects in cooperation with the National Institute for Public Health and Environment (RIVM). His work focuses on the quantitative risk assessment of consumer exposure to chemical substances. These substances enter our bodies through consumption of foods but also through products such as paint, cleaning products or cosmetics. Waldo is one of the main programmers of MCRA (Monte Carlo Risk Assessment). This is a web-based platform mainly used by researchers in life sciences. In MCRA, many statistical models are implemented to estimate the human exposure to substances in our daily food. The outcomes enable risk-managers to set out regulations to reduce risks. His other activities include advising and supporting research done by colleagues in agriculture or industry.