Geospatial Health 8(2), 2014, pp. 573-582

Optimizing denominator data estimation through a multimodel approach

Ward Bryssinckx[1], Els Ducheyne[1], Veerle Versteirt[1], Herwig Leirs[2], Guy Hendrickx[1]

[1]*Avia-GIS, Zoersel, Belgium;* [2]*Evolutionary Ecology Group, University of Antwerp, Antwerp, Belgium*

Abstract. To assess the risk of (zoonotic) disease transmission in developing countries, decision makers generally rely on distribution estimates of animals from survey records or projections of historical enumeration results. Given the high cost of large-scale surveys, the sample size is often restricted and the accuracy of estimates is therefore low, especially when spatial high-resolution is applied. This study explores possibilities of improving the accuracy of livestock distribution maps without additional samples using spatial modelling based on regression tree forest models, developed using subsets of the Uganda 2008 Livestock Census data, and several covariates. The accuracy of these spatial models as well as the accuracy of an ensemble of a spatial model and direct estimate was compared to direct estimates and "true" livestock figures based on the entire dataset. The new approach is shown to effectively increase the livestock estimate accuracy (median relative error decrease of 0.166-0.037 for total sample sizes of 80-1,600 animals, respectively). This outcome suggests that the accuracy levels obtained with direct estimates can indeed be achieved with lower sample sizes and the multimodel approach presented here, indicating a more efficient use of financial resources.

Keywords: spatial modelling, survey design, extensive livestock systems, multimodel models, Uganda.

Introduction

Animal distribution maps are essential for assessing disease transmission risk and providing estimates of poverty and nutritional needs (Kruska et al., 2003; IFPRI, 2010). In most developed countries this information is available through a mandatory livestock registration procedure (Augsburg, 1990). In developing countries, however, this information often is poor; often based on outdated distribution data or projections of enumeration results (Wanyoike et al., 2005).

Direct estimates of livestock numbers are made by multiplying the average number of livestock observed in sampled households by the total number of households. These data can be aggregated and mapped on different administrative unit levels as done for the Global Livestock Production and Health Atlas (GLiPHA) (Clements et al., 2002). This survey approach is very straightforward and easy to implement but it also has drawbacks. Being far from a full census, the results of are prone to sampling error. In case of a large sample size, sufficiently qualified staff is difficult to find and less well-trained enumerators and data entry personnel might lead to inaccuricies (FAO, 1996). In addition, survey costs can be high for large samples and the accuracy of livestock estimates in developing countries is therefore often restricted by available funding. In order to improve accuracy without increasing the number of samples, methods for improving data processing for livestock abundance maps are needed.

Spatial modelling is commonly adopted and spatial multiple regression has already been used for livestock estimates and this approach is extensively documented for the Gridded Livestock of the World (GLW) (Wint and Robinson, 2007). However, the assumptions made by this spatial modelling technique require statisticians to assess whether it is appropriate to apply them using given data. The livestock distribution maps from linear regression models offer continuous estimates of livestock abundance on a pixel by pixel basis, even in regions lacking samples. This leads to a false sense of accuracy since, in most cases, single pixel values cannot be used directly and data must be interpreted on regional basis. The extent at which individual pixel values must be aggregated in order to obtain reliable figures is unknown (Wint and Robinson, 2007).

Modelling can fill these gaps if adjusted to match total livestock numbers at the country level estimates, e.g. by the Food and Agriculture Organization (FAO). Another approach is to include more than one estimate in an ensemble of models using multimodel inference. This technique has been applied in a variety of

Corresponding author:
Ward Bryssinckx
Avia-GIS
Risschotlei 33, BE-2980 Zoersel, Belgium
Tel./Fax +32 3458-2979
E-mail: wbryssinckx@avia-gis.be

fields including species distribution modelling (Wintle et al., 2003). Bayesian inference and multimodel inference are most commonly used when combining multiple models (Link and Barker, 2006). Both these approaches require weighting of the different models to represent how well each model approximates the "true values". The methodologies also differ with respect to the selection criterion used to assess which model fits observed data best. Both Akaike's information criterion (AIC) and a Bayesian information criterion (BIC) have been used for this purpose (Wintle et al., 2003; Burnham and Anderson, 2004; Link and Barker, 2006). The latter two authors suggest that assigning weights should be based on model evaluation using different datasets. Regardless which weight assignment method is applied, the models included in the ensemble should have sufficient informative value in order to contribute to the added value of the ensemble. As livestock distribution data in countries with predominantly extensive livestock systems are sparse and often outdated, prior knowledge for a BIC is in most cases inaccurate. The ensemble approach proposed in this study includes direct estimate of livestock numbers. When looking at this direct estimate as one of the models in the ensemble, the AIC would assign most of the weight to the direct estimate as it perfectly fits the training data. Therefore, AIC is not suitable for weight allocation either. Hence, with the overall aim of increasing accuracy in extensive systems where the number of survey samples is limited, another ensemble approach is proposed for processing georeferenced livestock counts. Direct estimates were averaged with regression tree forests (RTF), or random forest results, for which use is not restricted to continuous predictor variables or unimodal data. The aim of this work is to assess if this data processing technique is able to increase the accuracy of livestock abundance reports based on agricultural survey data.

Since the multimodel approach was foreseen to be part of an unsupervised algorithm based on modelling techniques less prone to collinearity and distributional assumptions, linear regression was avoided and instead random forests adopted to estimate denominator data. A random forest consists of multiple regression trees, which partition data according to a series of splits in one or more dimensions. This results in a structure which can be used to classify a data point for which the value is not known. The value to be addressed at the data point is the mean of values within the class which was found by answering the binary questions at each split of the regression tree.

With the aim in mind to assess if this data processing technique would be able to increase the accuracy of livestock abundance reports based on agricultural survey data, direct estimates were averaged with random forest results, for which the use is not restricted to continuous predictor variables or unimodal data.

Materials and methods

Study site

Uganda is a middle-sized country in the eastern part of Africa (mostly between 4° N and 2° S, and 29° E and 35° E) surrounded by South Sudan, Kenya, Tanzania, Rwanda and the Democratic Republic of the Congo. The total area is 241,139 km^2, 18% of which is covered by fresh-water lakes. Although this could be an ideal water source for livestock, the major lakes (Lake Victoria, Lake Wamala, Lake Albert, Lake George and the Kyoga/Kwania lake complex) that make it up is more of interest for the fishery industry (Ibale, 1998). While the average altitude is about 1,100 metres above mean sea level (with Mount Stanley the highest peak at 5,113 m), lower altitudes are found near the South Sudanese plain in the north. The relatively high altitude tempers the tropical climate at 16-26 °C between April and November but 30 °C is generally exceeded between December and March. While rainfall is most abundant in the South (>2,100 mm), where areas are well covered by vegetation, it decreases towards the Northeast (500 mm) resulting in savannas and dry plains.

In rural communities, agricultural production and livestock abundance do not only indicate food security, but also reflect current and future economic security as replacement of livestock owned by a household is believed to be home-bred thus limiting the extra need for financial capital to sustain or increase herd sizes. In contrast, without previous livestock ownership, start-up is difficult requiring spending savings (Upton, 2004). In the eastern, northern and western region, the majority of households live in rural communities in which 72% of all households owns at least one kind of livestock. In the central region, urban settlements are more numerous and only 56% of all households owns livestock (MAAIF, 2009). However, agriculture also plays a major role in urban areas, as food prices determine the livelihood of people living there.

Livestock-rearing in some parts of northern Uganda is complicated by the presence of armed rebel movements forcing people to stay in refugee settlements with uncertainty of feeding an the general economy. Although relative safety has returned, there is still a long way to restore livestock numbers to previous numbers (USAID, 2008). Uganda is currently subdivided in 111 districts, an increase compared to the period between 2007 and 2009 when only 80 districts existed (mean surface 3,000 km^2). Districts are further divided into counties (mean surface 1,500 km^2), subcounties (mean surface 200 km^2) and parishes (mean surface 45 km^2) containing several villages.

Livestock data

In 2008, the Animal Industry and Fisheries (MAAIF) division of the Ministry of Agriculture, implemented an extensive livestock survey with technical support for data collection, processing and reporting from the Uganda Bureau of Statistics (UBOS). A two-stage sampling approach was applied with sub-counties were randomly selected from each district at the first stage. At the second stage, at least 50 enumeration areas (EAs) were sampled from each selected sub-county. One EA includes 200 households. The same enumeration area sampling frame as developed for the 2002 Population and Housing Census was applied. In total 8,870 EAs were enumerated. Previous livestock counts date from 2002, i.e. the Population and Housing Census (UBOS, 2002) and 2005, i.e. the Uganda National Household Survey (UBOS, 2007).

Modelling livestock distribution

Obtaining direct estimates of cattle numbers through aggregating sampled data on a specified administrative unit level is straightforward. It is also self-evident that the accuracy of direct estimates will be higher for a larger number of samples. Reporting denominator data on a lower administrative unit level (smaller area size) will diminish the number of EAs that belong to one unit and therefore rato of estimate sensitivity to sampling error will increase. When validating a statistical model using resampling procedures, the estimated values for training data entries will tend to regress towards an average value of observed cases in similar conditions. This generalization is used in the proposed methodology to lower the degree of under- and over-estimation of cattle numbers. Given the restrictions associated with unsupervised modelling, e.g. using linear regression techniques, a random forest approach was followed. For each forest 500 trees were grown. At each split one third of the predictor variables were randomly sampled as candidates. The minimum size of terminal nodes was set to five observations. Before the models were trained, a logarithmic transformation was applied on the denominator data in order to have similar residuals for different livestock numbers. Before error assessment, back transformation was done.

Predictor variables

Distance to water in km was deduced from the consolidated VMap0 river-surface water body network data (NIMA, 1997) and used as proxy for access to drinking water for herds (Luke, 1987). A preliminary analysis indicated that the majority of large herds were located at moderate altitudes (1,000-1,500 m) with low slope values (0°-2°). Given that denominator data responded differently for increasing altitude and slope values, both variables were retained as candidate predictor variables.

The total human population estimates, as prepared by the FAO for the World Bank's rural development strategy review, was included since it was shown that there is a positive correlation between human and livestock population densities (Lapar and Jabbar, 2003). However, for urban areas with high population densities, industrialisation might offer job opportunities other than agricultural activities such as livestock rearing (Iruonagbe, 2009). As surrogate for this urban character, "night-time lights" (NOAA National Geophysical Data Center, 2003) was used (Amaral et al., 2006). Accessibility is another auxiliary variable related to economic activities, which gives the estimated travel time (in min) needed to reach a major city, i.e. a city with 50,000 or more people in the year 2000 (JRC Global Environmental Monitoring Unit, 2008).

Long-term, average monthly temperature and precipitation values were derived from the WorldClim dataset (Hijmans et al., 2005). For all continuous variables, the mean value was assigned to the administrative units. Finally, the GLC 2000 1-km global land cover (JRC Land Resource Management Unit, 2000) was used for land cover classification. The majority class was assigned to the administrative unit resulting in four predominant land cover types: crop/forest, croplands, forest and shrublands. A suitability mask was used to exclude unpopulated areas such as water-bodies, game and nature reserves from the predictor variable layers.

The targetted added value of applying random forest consists in averaging under- and over-estimates of cattle numbers. This generalisation of denominator data will not always be beneficial for the accuracy of estimates as is the case when the number of animals really deviates from what one would expect based on the assessment of environmental parameters. A wrong estimate of cattle numbers can arise from an incomplete set of predictor variables. Missing predictor variables are very likely as herding of cattle or other domestic animals remains the choice of the individuals in the human population. Their decision to rear animals depends on many variables, which may not all be covered by available datasets.

Because direct estimates do capture local differences in livestock abundance, the performance of a multimodel approach was assessed as well. The most notable difference with existing multimodel approaches is that only two models were considered: (i) a direct estimate of livestock numbers where inaccuracies mostly result from sampling error; and (ii) a regression tree forest which generalises livestock abundance for subsets of similar administrative units. Both these models are averaged using uninformative weights (unweighted average). A multimodel approach is also more likely to represent a consensus opinion of the two approaches, the closest of which to what would have been observed if a total census was performed, would not be known (Millington and Perry, 2011).

Sample subsets

In order to test the robustness of the multimodel approach, 250 survey simulations were tested with different sample subsets for each run. For reasons of simplicity, tests were restricted to a spatially random sampling strategy. No other strategies such as stratified or clustered sampling were considered. Because the sample density of the Uganda 2008 National Livestock Census is not homogeneous throughout the entire country, it is not sufficient to take a random sample from the list of visited EAs. Instead, locations within the study area had to be selected randomly and the nearest visited EA sought to approximate a spatially random sample. As long as the number of samples in the subset remains small as compared to the total dataset, it was assumed that samples were not depleted locally and randomness was assured. The vast number of samples taken during the Uganda 2008 Livestock Census enabled generating input datasets for all repetitions as well as comparing results from a series of increasing sample sizes. In this study, total sample sizes of 80, 240, 400, 560, 720, 880, 1,040, 1,200, 1,360, 1,520 and 1,600 EAs were considered to evaluate the impact of varying sample size. The random selection of samples was automated in the R statistical computing and graphics environment (R Development Core Team, 2012).

Accuracy assessment

The sample sizes to be tested were chosen so that survey data subsets, taken in each repetition to estimate livestock numbers, only represented a small part of the total number of entries. The small subset was regarded as being the result of a simulated, small-scale survey, while the total set was regarded as representing the entire population. When assessing the error of direct estimates or random forest outcomes, the difference between the estimate and aggregated survey entries summarised per administrative unit was calculated. The enormous extent of the Uganda 2008 National Livestock Census permits the testing of a large range of sample sizes, which are common among livestock surveys in other countries or field studies.

Before averaging RTF results and direct estimates, the performance of the direct estimates and regression tree forests was compared by calculating a standardised error measure. For each repetition and district, a relative error was calculated for both RTF results and direct estimates. Each pair of relative error values was standardised so they summed up to one. Per abundance class, a mean value of the error measure for direct estimates was calculated over the repetitions and districts for each sample size. A value of 0.50 means that both approaches estimate the denominator data equally well, while a value higher than 0.50 means that RTF results outperform direct estimates. The plot must thus be interpreted as to how well the number of cattle per household is assessed by RTF results compared to the direct estimates. Overall accuracy and accuracy per livestock abundance class was assessed. Livestock abundance classes were defined using thre different approaches (Fig. 1): k-means classification (Hall and Ball, 1965), quantiles and natural breaks (Bivand et al., 2008). Out of these, the k-means classification method results in the most balanced distribution of training data among an arbitrarily chosen number of three classes. For k = 3 clusters, the global k-means clustering optimum is reached within 10 repetitions. The resulting class intervals are defined in Table 1.

Table 1. Livestock abundance classification. Class intervals and frequency distribution of counties with a small, medium-sized and large average herd size.

	Minimum	Maximum	Frequency
Small herds	0.0060	2.7124	120
Medium-sized herds	2.7124	9.1090	26
Large herds	9.1090	29.9219	15

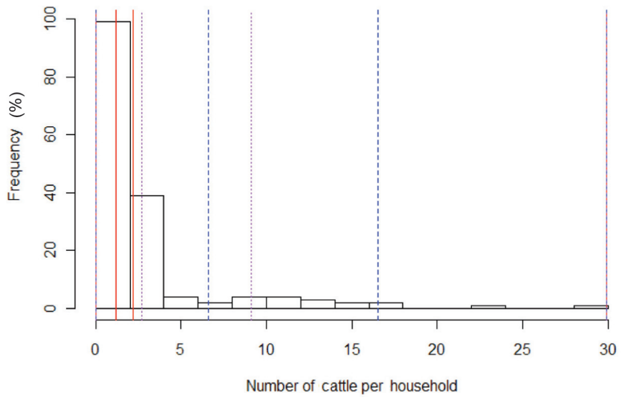

Fig. 1. Livestock abundance classification. Classification of denominator data into three categories using quantiles (red), natural breaks (blue) and k-means (purple) classification methods.

Table 2. Statistical significance testing of the land cover classes. The resulting P-values of the Mann Whitney U test show that no significant difference in accuracy improvements was found among different land cover classes.

	1st quantile	2nd quantile (median)	3rd quantile
Crop forest	0.8789	0.8411	0.8537
Croplands	0.9390	0.7376	0.8557
Forest	0.5747	0.9336	0.9424
Shrublands	0.4977	0.6956	0.6202

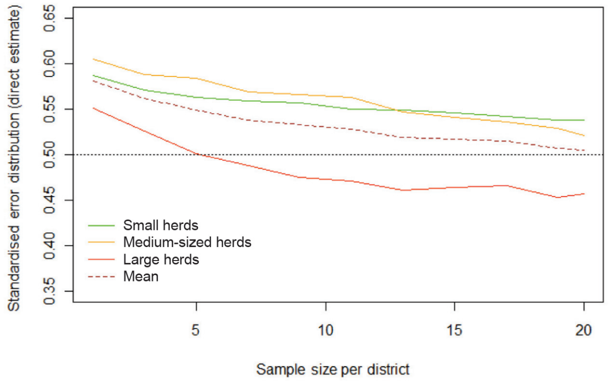

Fig. 2. Comparison of relative error values between model predictions and aggregated denominator data. The complete set of simulation results was split into districts with small herds, with medium sized herds and with large herds. Although a higher number of samples improves model performance, aggregated denominator data (direct estimate) benefits more of an increased sample size (hence the descending lines on the graph). For large sample sizes, the mean error distribution approaches 0.50 suggesting that the direct estimate method and model predictions perform equally well.

A Wilcoxon matched pairs signed rank test was performed to assess whether pairwise comparison showed the unweighted approach resulting in a location shift to lower relative error values compared to the direct estimate. The statistical difference in accuracy among different dominating land cover classes per districts was determined using a Mann Whitney U test (R Development Core Team, 2012). Accuracy improvement was expressed as the relative error of the unweighted mean approach divided by the relative error of the direct estimate. The first, second and third quantiles were tested. To evaluate land cover class differences no distinction was made between different herd sizes in order to retain sufficient training data. The null hypothesis states that there is no significant difference between accuracy improvements within different land cover classes.

Results

The relative performance of direct estimates and RTF is shown in Fig. 2. The lower the sample size, the better the RTF approach estimates the denominator data compared to the direct estimates. For one sample per district, relative error values of aggregated sample data are 1.58, 1.47 and 1.16 times larger than model outputs for the low, medium and high abundance class, respectively. For a number of five samples per district, performance of both approaches was comparable for large herds. While relative error values of aggregated sample data compared to model outputs kept decreasing for larger sample sizes, the rate of decay diminishes. For a number of 20 sampled EAs per district, the mean relative error of aggregated sample data equaled that of model outputs.

Because of multiple districts within one denominator data class and a multitude of repetitions, the accuracy is described by a distribution. To get a clearer view on relative error differences between the applied methodologies, a minimum value, a maximum value and all three quartiles were computed for each repetition. The median relative error values for 250 repetitions are given for each abundance class in Fig. 3. For districts with small, medium-sized and large herds, the mean differences over all tested sample sizes between median relative error values of direct estimates and unweighted averages were 5.3%, 11.4% and 6.9%, respectively. Differences between direct estimates and model results were 5.1%, 9.0% and -2.8%, respectively.

For small sample sizes, aggregated denominator data generally performed far worse than model outputs.

Therefore, the model outputs showed smaller relative error values than both aggregated denominator data and the unweighted mean. When the sample size was increased, the model output did not outperform aggregated denominator data consistently and unweighted mean values were better than both aggregated and modelled values. For large herds, this happened when the sample size exceeded one sample per district, for medium-sized herds more than two samples per district were required and small herd districts generally needed more than three samples per district to let unweighted mean values outperform the other approaches.

In districts with high livestock abundance, the spatial model only performed better than the direct estimate when the sample sizes were low. From three samples per district and onward, the model was outperformed by the direct estimate. However, the model still contributed to making more accurate predictions when both results were combined into a joint livestock abundance estimate (unweighted average). Also for administrative units smaller than districts, relative error comparisons showed that an unweighted mean value of aggregated and predicted denominator data generally performs better than the direct estimate (not shown in the figures).

When combining all livestock abundance classes, the mean median relative error difference between direct estimates and RTF results equaled 5.3%. For the difference between direct estimates and unweighted averages, the mean accuracy increase equaled 6.5% with a maximum of 16.6% for low sample sizes and a minimum of 3.7% for high sample sizes (Fig. 4). When calculating these mean differences, tests for all sample sizes were included. Both for district level as well as county administrative unit ones, the null-hypothesis of independence of estimates was rejected at the 99% level of significance for all tested sample sizes based on Wilcoxon matched pairs signed rank tests. For individual land cover classes, similar negligible P-values were observed.

When the administrative units were categorised into subsets according to their land cover class, similar P-values were obtained. Within each class the relative errors of the unweighted mean approach were also divided by the relative errors of the direct estimate, to get an index of accuracy improvement (Figs. 5 and 6). In each land cover class, only few districts performed worse (i.e. those where the relative error of the unweighted mean was larger than the relative error of the direct estimate).

For low sample sizes (n = 80, i.e. 1 sample per district), a similar small number of districts experienced an adverse effect of the unweighted mean approach. In

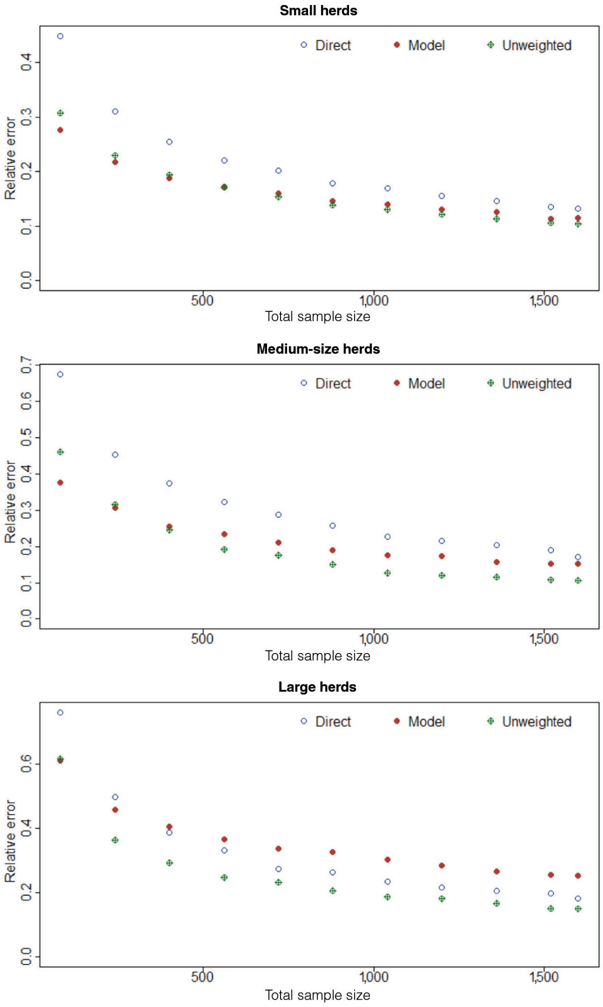

Fig. 3. Accuracy increase (districts split according to average herd size). Results given at the district level with separate plots for different herd sizes. Median relative error values for aggregated sample data (blue), model outputs (brown) and unweighted model-aggregate data (green). In general, the unweighted model-aggregate data performs best, except for very small samples where aggregated sample data generally performs far worse than the model suggesting the model on its own as the best approach.

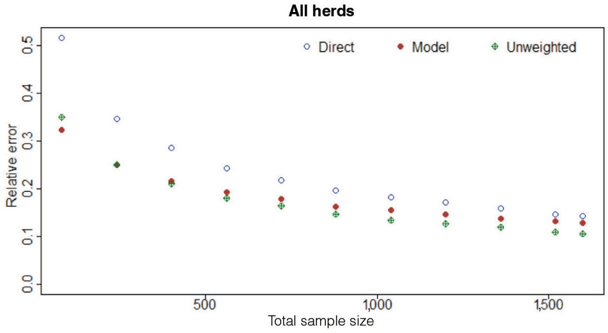

Fig. 4. Accuracy increase (all districts). Median relative error values for livestock number estimates including districts of all livestock abundance classes.

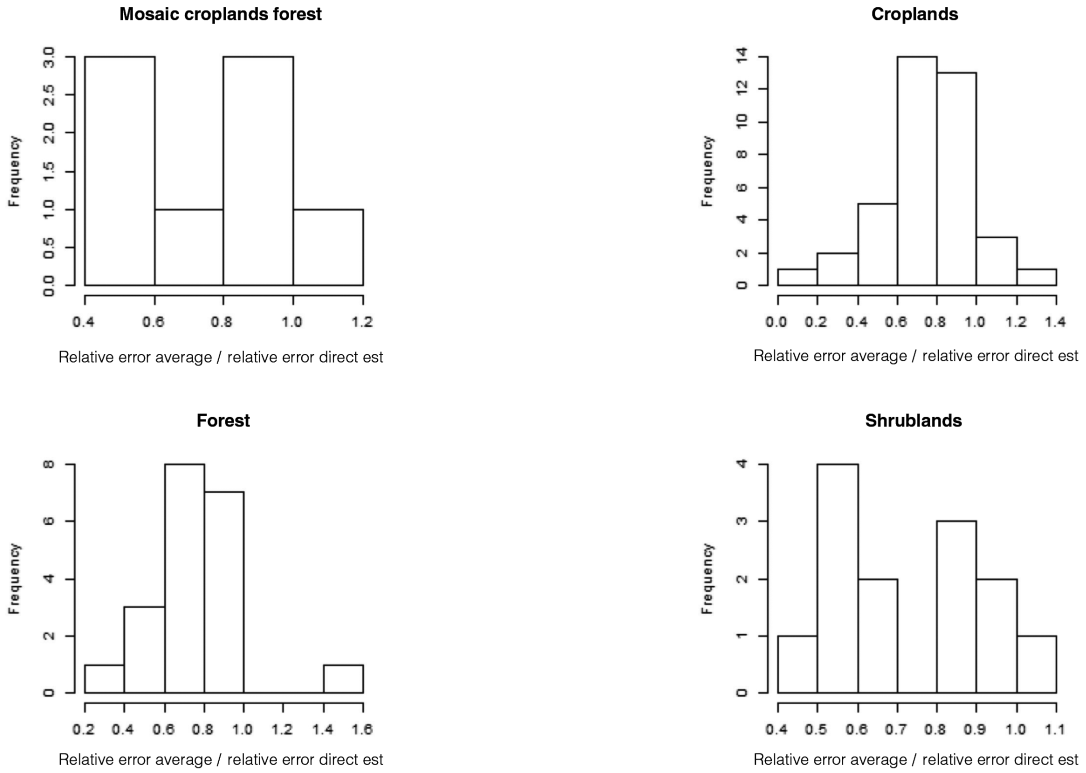

Fig. 5. Median accuracy improvement indices per land cover class on district level. Values larger than 1 indicate an accuracy decrease. As can be seen clearly, the vast majority of districts shows a small to large improvement. The sample size used here was 1,040 enumeration areas.

the croplands land cover class, one district (Kampala) showed a relative error for the unweighted mean approach, which was 1.8 times larger than the relative error of the direct estimate. The Kampala district, which is conterminous with Uganda's capital city, only holds 0.085 cattle per household which is by far the smallest number compared to other districts.

When the accuracy improvements of the unweighted mean approach over the direct estimate were compared for the sample size of 1,040 using the Mann Whitney U test (Mann and Whitney, 1947), the p-values found indicate that there is no significant difference in accuracy between the different land cover classes (Table 2). Also for the other sample sizes, no significant difference among the land cover classes was observed. The same phenomenon was seen at the county level. For Bwamba county in the Bindibungyo district, the mean number of cattle per household was 0.0225 (compared to 2.9510 animals per household for the whole of Uganda). Even a small overestimation of this number by the spatial model could, in this case, lead to a very large relative error.

The populated area of Bwamba measures 181.5 km^2 which makes it one of the 10 smallest, populated areas within a county. In such small areas, the risk of only encountering uncommonly low or high livestock numbers among the local population is larger as variation among livestock numbers becomes more likely to vary with larger distances. Such distances cannot be found between households within small administrative units such as Bwamba. In total 3,643 households were sampled in Bwamba (as compared to a median number of 4,623 households per county and a minimum number of 202). While a direct estimate can estimate livestock numbers accurately based on such a large sample size, a spatial model would experience difficulties by doing so as the number of cases with extreme livestock numbers is limited among the training data.

Discussion

Livestock numbers in Uganda have more than doubled over the last three livestock surveys, carried out in 2002, 2005 and 2008. As an increase of 6.2 million animals (as compared to 5.2 million in 2002) within 6 years is rather unlikely, this may be partly attributed to the use of different sampling frames and the huge sample size in 2008 (15.1% of the total number of households were visited). Re-stocking under the national livestock productivity improvement project, livestock

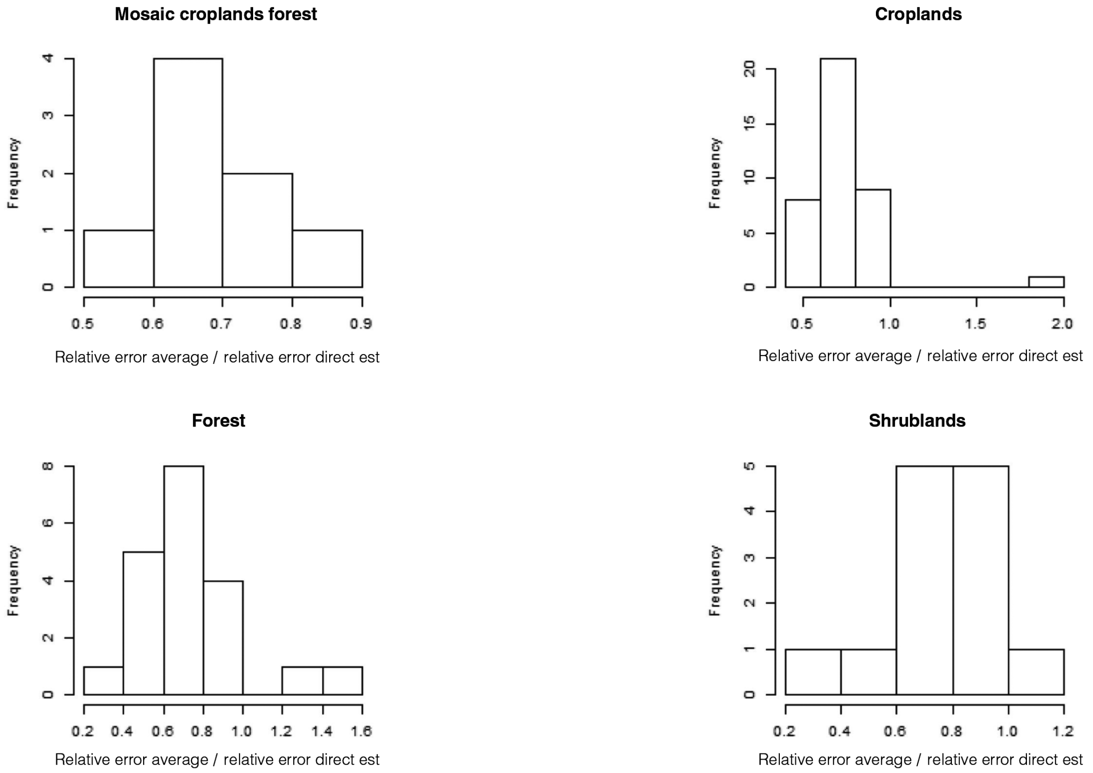

Fig. 6. Median accuracy improvement indices per land cover class on district level. Values larger than 1 indicate an accuracy decrease. The sample size used here was 80 enumeration areas (one sample per district).

becoming a lucrative enterprise due to an increasing demand for beef, is another possible reason, as are strategies implemented by MAAIF, such as carrying out effective disease control and increasing acreage of land utilised for cattle rearing (MAAIF, 2009).

Based on the 2008 livestock survey data, modelling denominator data appeared to be an effective way of improving the accuracy of estimates compared to direct estimates. However, when the results were subdivided by livestock abundance, it became clear that in the few districts with large herds, prediction performance was worse compared to the direct estimate when sample sizes were small (<400 EAs for all 80 districts). Moreover, when only a small livestock survey is available, and no extensive validation dataset (such as in this testing framework) is at hand, there is uncertainty about which model results are closest to the truth. As Wintle et al. (2003) noted, there is considerable risk in ignoring alternative arguments (i.e. outcomes from other models) when selecting one single model. A more conservative approach would be to consider any of the competing models by calculating an average estimate and assign weights according to the degree of belief. In this case, the degree of belief depends on the number of administrative units with a similar number of livestock compared to the number of administrative units in other livestock abundance classes. As the true number of livestock is not known (sampling error may cause a vast increase or decrease of the direct livestock number estimate), no weights were given to either the direct or the spatial model estimate.

When an ensemble estimate was made by computing the unweighted mean of a direct estimate and spatial model prediction, the relative error values were further reduced and a general accuracy improvement was obtained over the entire range of tested sample and herd sizes. This shows the proposed approach to be a robust way of improving denominator data estimates without increasing sample size. This is accomplished through averaging under- and over-estimations, and through the introduction of data from outside the administrative unit of interest.

Observed accuracy differences are the result of an unbalanced training dataset with many administrative units with small cattle numbers per household and only few administrative units where households own a relatively large number of cattle. When using the proposed methodology, accuracy will therefore be highest for administrative units in the livestock abundance class, which is most common among the training data. As this plays to the advantage of those who will use the livestock distribution maps, there is no need to balance the training dataset by removing some of the administrative units of the most common livestock abundance class. This would only diminish the amount of training data greatly and they are usually already sparse.

Because a very comprehensive sample size was available for validating the proposed methodology, large sample sizes could be tested. Therefore, the results are considered to be applicable in many countries as sample sizes are usually much smaller in most national livestock surveys covering mixed-farming systems. Which additional predictor variables should be added to the model may depend on country-specific situations and might require specific adaptation. While the spatial level on which denominator data is evaluated was crucial for the positive outcome of ameliorating denominator data estimates by applying spatial interpolation (Bryssinckx et al., 2012), this is much less of a concern when averaging direct estimates and spatial model outcomes. The reason is mainly because of the dependency on very local data for the spatial interpolation methodology, while the unweighted mean approach depends on data from the entire study area through a regression tree based model. In addition, the use of fixed weights given to the model predictions and aggregated survey results provides a more robust way of processing survey data compared to spatial interpolation, where spatial scale and geometry of administrative units have a significant impact on how weights are distributed.

Sampling more EAs resulted in an increased performance of the spatial model. Therefore the accuracy of the unweighted mean approach should also improve with the size of the sample. The greatest challenge for the spatial model in this approach seems to lay in estimating extremely low denominator data values. Therefore, it is important to include predictor variables capable of differentiating between livestock numbers when the abundance is low.

Conclusion

Using an ensemble of aggregated livestock numbers and spatial model outputs as a joint denominator data estimate proved to be a viable method for improving accuracy without raising survey costs. While accuracy improvements were clear on each administrative unit level, maximum relative errors of the unweighted mean approach were higher than those of the direct estimates when data was aggregated on an administrative unit level smaller than districts. This was due to extremely small denominator data values which the spatial model was not able to predict. In those few cases, the absolute error was small. This also shows that the proposed method should not be applied when the objective is to detect areas with extremely low or high livestock numbers. Instead, the method results in more robust estimates by using data of similar environments. This also means that aggregated denominator data from the administrative unit's sample will, to some extent, regress towards the mean of denominator data in environmentally corresponding parts of the study area. Although the performance of spatial models in other settings (areas) is difficult to assess without further research, this study shows how livestock survey results can be improved without additional data needs (except for spatial covariate data which is easily accessible through various sources at no additional cost). This also implies that the cost-efficiency can be enhanced due to reduced sample size requirements for similar accuracy levels. How sample size requirements for this new method can be assessed in detail is subject for further work.

Acknowledgements

The authors wish to thank the Ministry for Agriculture, Animal Industry and Fisheries (MAAIF) and the Uganda Bureau of Statistics (UBOS) for making available the Uganda 2008 National Livestock Census data and Albert Mugenyi for his valuable assistance and contributions. The authors also gratefully acknowledge the funding support from IWT and the ICONZ project.

References

Amaral S, Monteiro V, Camara G, Quintanilha JA, 2006. DMSP/OLS night-time light imagery for urban population estimates in the Brazilian Amazon. Int J Remote Sens 27, 855-870.

Augsburg JK, 1990. The benefits of animal identification for food safety. J Anim Sci 68, 880-883.

Bivand RS, Pebesma EJ, Gómez-Rubio V, 2008. Applied spatial data analysis with R. Use R! New York: Springer (second edition), 374 pp.

Bryssinckx W, Ducheyne E, Muhwezi B, Godfrey S, Mintiens K, Leirs H, Hendrickx G, 2012. Improving the accuracy of livestock distribution estimates through spatial interpolation. Geospat Health 7, 101-109.

Burnham KP, Anderson DR, 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33, 261-304.

Clements ACA, Pfeiffer DU, Otte MJ, Morteo K, Chen L, 2002. A global production and health atlas (GLiPHA) for interactive presentation, integration and analysis of livestock data. Prev Vet Med 56, 19-32.

FAO, 1996. Conducting agricultural censuses and surveys. Rome: Food and Agriculture Organization.

Hall DJ, Ball GB, 1965. ISODATA: a novel method of data analysis and pattern classification. Technical report, Stanford Research Institute, California. Available at: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=AD0699616 (accessed on February 2014).

Hijmans RJ, Cameron JL, Parra PGJ, Jarvis A, 2005. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25, 1965-1978.

Ibale RDW, 1998. Towards an appropriate management regime for the fisheries resources of Uganda, Entebbe. Ministry of Agriculture, Animal Inustry and Fisheries.

IFPRI, 2010. Livestock development planning in Uganda: identification of areas of opportunity and challenge. Kampala: International Food Policy Research Institute.

Iruonagbe TC, 2009. Rural-urban migration and agricultural development in Nigeria. Arts Soc Sci Int J 1, 28-49.

JRC Global Environmental Monitoring Unit, 2008. Travel time. Available at: http://bioval.jrc.ec.europa.eu/products/gam/index.htm (accessed on February 2014).

JRC Land Resource Management Unit, 2000. Global Land Cover 2000 Project. Available at http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php (accessed on February 2014).

Kruska RL, Reid RS, Thornton PK, Henninger N, Kristjanson PM, 2003. Mapping livestock-oriented agricultural production systems for the developing world. Agr Syst 77, 39-63.

Lapar MLA, Jabbar MA, 2003. A GIS-based characterisation of livestock and feed resources in the humid and sub-humid zones in fove countries in South-East Asia. Nairobi: International Livestock Research Institute. CASREN paper no. 2, 72 pp.

Link WA, Barker RJ, 2006. Model weights and the foundations of multimodel inference. Ecology 87, 2626-2635.

Luke GJ, 1987. Consumption of water by livestock. Technical report. Australia: Department of Agriculture Western Australia. Available at: http://www.agwestinternational.wa.gov.au/objtwr/imported_assets/content/aap/sl/nut/tr060.pdf (accessed on February 2014).

MAAIF, 2009. The national livestock census report 2008. Entebbe: Ministry of Agriculture, Animal Industry and Fisheries. Available at: http://www.agriculture.go.ug/userfiles/National%20Livestock%20Census%20Report%202009.pdf (accessed on February 2014).

Mann HB, Whitney DR, 1947. On a test of whether one of two random variables is stochastically larger that the other. Ann Math Stat 18, 1-164.

Millington JDA, Perry GLW, 2011. Multi-model inference in biogeography. Geography Compass 5, 448-463.

NIMA, 1997. VMAP_1V10 - Vector Map Level 0 (digital chart of the world). Fairfax: National Imagery and Mapping Agency Available at: www.mapability.com/info/vmap0_download.html (accessed on February 2014).

NOAA National Geophysical Data Center, 2003. Nighttime lights of the world. Available at: http://sabr.ngdc.noaa.gov/ntl/ (accessed on December 2011).

R Development Core Team, 2012. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

UBOS, 2002. 2002 population and housing census. Provisional results. Entebbe: Government of Uganda. Available at: http://www.ubos.org/onlinefiles/uploads/ubos/pdf%20documents/2002%20Census%20Final%20Reportdoc.pdf (accessed on February 2014).

UBOS, 2007. Uganda national household survey 2005/2006. Report on the agricultural module. Entebbe: Uganda Bureau of Statistics. Available at: http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1181743055198/3877319-1328111100912/UNHS.2005-06.Agriculture.Report.FINAL.pdf (Accessed on February 2014).

Upton M, 2004. The role of livestock in economic development and poverty reduction. PPLPI Working Papers. Rome: Food and Agriculture Organization of the United Nations. Available at: http://www.fao.org/ag/againfo/programmes/en/pplpi/docarc/wp10.pdf (accessed on February 2014).

USAID, 2008. Livestock health services in northern Uganda baseline survey. Washington DC: United States Agency for International Development. Available at: http://pdf.usaid.gov/pdf_docs/PNADP078.pdf (accessed on February 2014).

Wanyoike F, Nyangaga J, Kariuki E, Mwangi DM, Wokabi A, Kembe M, Staal S, 2005. The Kenyan cattle population: the need for better estimation methods. Nairobi: Smallholder Dairy Project. Available at: http://cgspace.cgiar.org/handle/10568/2214 (accessed on February 2014).

Wint W, Robinson T, 2007. Gridded livestock of the world. Rome: Food and Agriculture Organization of the United Nations.

Wintle BA, McCarthy MA, Volinsky CT, Kavanagh RP, 2003. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conserv Biol 17, 1579-1590.