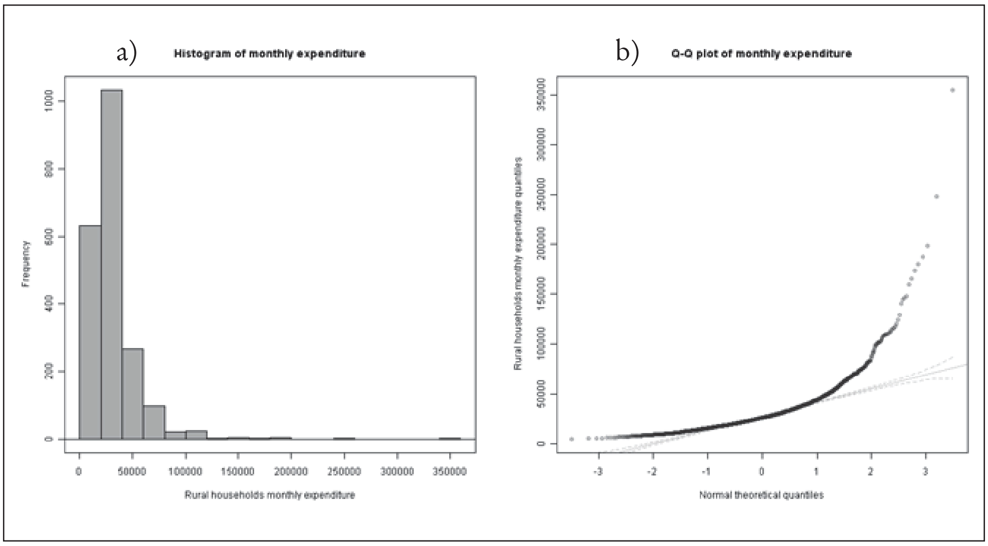
Results

TRANSFORMATION OF THE EXPENDITURE DATA

A simple histogram and q-q plot[5] of the aggregated rural monthly adult equivalent expenditure values confirmed their distribution to be far from normal (Figure 5).

Figure 5 Distribution of the rural monthly adult equivalent expenditure at 0.01 degree resolution (a), and q-q plot of quantiles against a theoretical, normal distribution (b).



The Box-Cox power transform (Box and Cox 1964) was used to normalise the distribution, which uses a power parameter, λ, and has the general form:

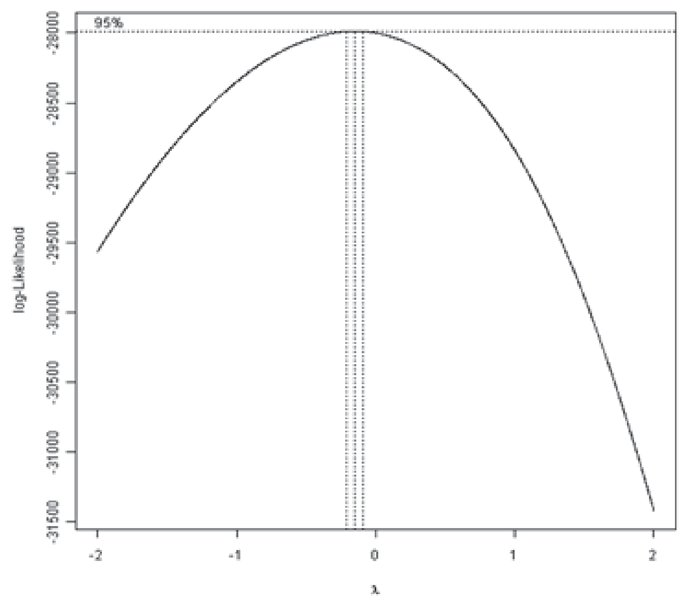
$$y^{bct} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & f \quad \lambda \neq 0 \\ \ln(y) & f \quad \lambda = 0 \end{cases}$$

A simple procedure in R computes and plots log-likelihoods for λ with 95 percent confidence limits (Figure 6). If zero lies within the confidence limits then ln(*y*) would be the more appropriate transform. In this case, λ was -0.151 and zero was not contained within its confidence limits, suggesting the Box-Cox transform to be the more appropriate transformation.

---

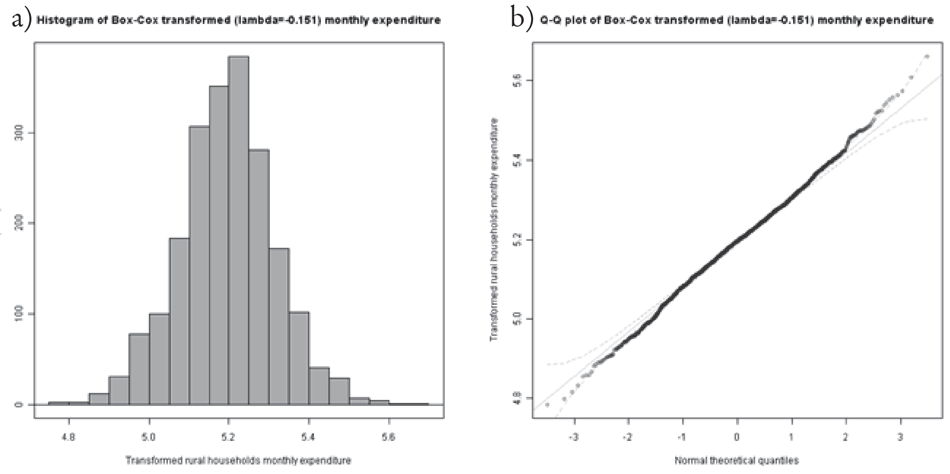
[5] a Q-Q plot (‘Q’ stands for quantile) is a probability plot, a kind of graphical method for comparing two probability distributions, by plotting their quantiles against each other. Here we compare the household data distribution against a normal distribution.

Figure 6. Log-likelihood for **λ** in the Box-Cox transformation at 0.01 degrees.



The histogram and q-q plot (Figure 7) of these transformed expenditure data show the distribution to be much more normal.

Figure 7. Distribution of the transformed rural monthly adult equivalent expenditure at 0.01 degree resolution (a), and q-q plot of quantiles of transformed data against a theoretical, normal distribution (b).


a) Histogram of Box-Cox transformed (lambda=-0.151) monthly expenditure

b) Q-Q plot of Box-Cox transformed (lambda=-0.151) monthly expenditure

The inverse transform was applied to obtain the resulting predicted welfare in the original units:

$$y = \begin{cases} (\lambda y^{bct} + 1)^{1/\lambda} & f \quad \lambda \neq 0 \\ e^{y^{bct}} & f \quad \lambda = 0 \end{cases}$$

Table 2 shows the λ values for each resolution along with the number of pixels that contained household data. At resolutions coarser than 0.75 degrees there are too few pixels containing data to fit reliable regression models.

Table 2. Pixel counts and λ values for the Box-Cox transformation at each spatial resolution (in decimal degrees).

Cell size	0.01	0.02	0.03	0.05	0.10	0.15	0.20
# of pixels	2 088	1 279	1 086	813	539	399	280
λ	-0.151	-0.074	-0.027	0.035	0.104	0.206	0.300
Cell size	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**	**0.75**
# of pixels	206	167	120	103	82	75	36
λ	0.359	0.325	0.250	0.472	0.113	0.465	0.898

VARIABLE SELECTION AND DESCRIPTION

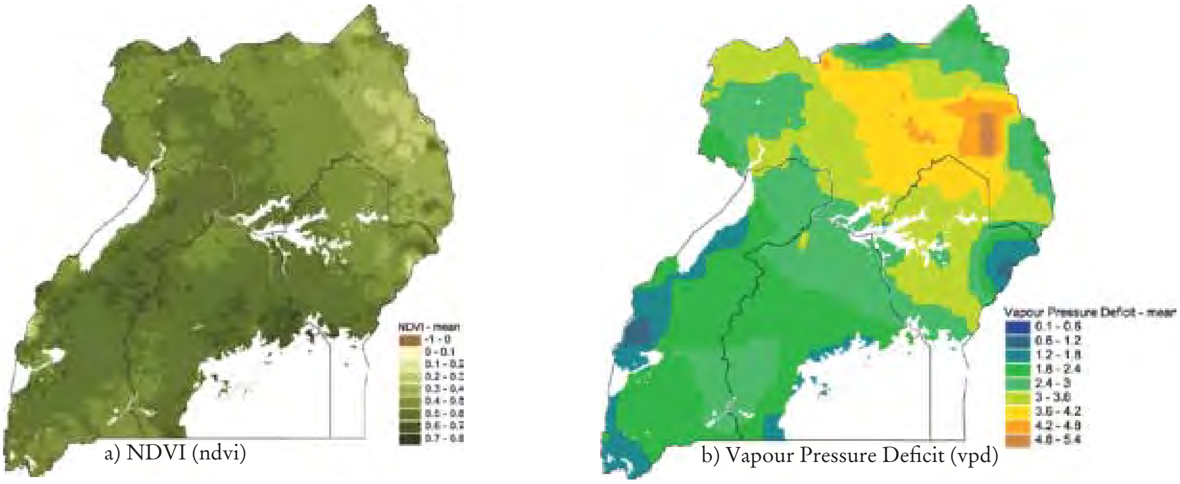
Based on the outcomes of previous studies, preliminary analyses and data availability the following were selected as independent, predictor variables: (i) mean annual NDVI (ndvi); (ii) mean annual Vapour Pressure Deficit (vpd); (iii) slope (slp); (iv) goat density (goat); (v) cattle density (cattle); (vi) travel time to markets (dist); and (vii) population density (grump).

These seven variables, at 0.01 degrees resolution, are shown in Figure 8. NDVI and VPD show variation in climate from the more humid central and southern regions to the arid northern and eastern regions. Goat densities are highest in the northeastern and southwest regions whereas cattle are found in a broad band spanning from the southwest to the northeast; the so-called ‘cattle coridoor’. Population density (grump) is higher, and access to markets (dist) better, in the central and southern regions than elsewhere. Finally, slp reflects well the mountainous terrain in the eastern and southern regions.

Table 3 shows a correlation matrix of the dependent (ybct) and independent variables. Firstly this demonstrates that there are no major collinearities among the independent variables. Secondly it shows that two of the independent variables, NDVI (+ve) and VPD (-ve), have stronger correlations with ybct than the other variables, goat (-ve), cattle (+ve), slp (+ve), grump (+ve), dist (-ve). The signs are broadly as expected, with the exception of slp, although the correlation is very weak, and possibly goat density (goats are predominant in the less wealthy pastoral and agro-pastoral areas of the northeast of Uganda). Table 4 gives a numerical summary of each variable, showing the skewed distributions of several of them.

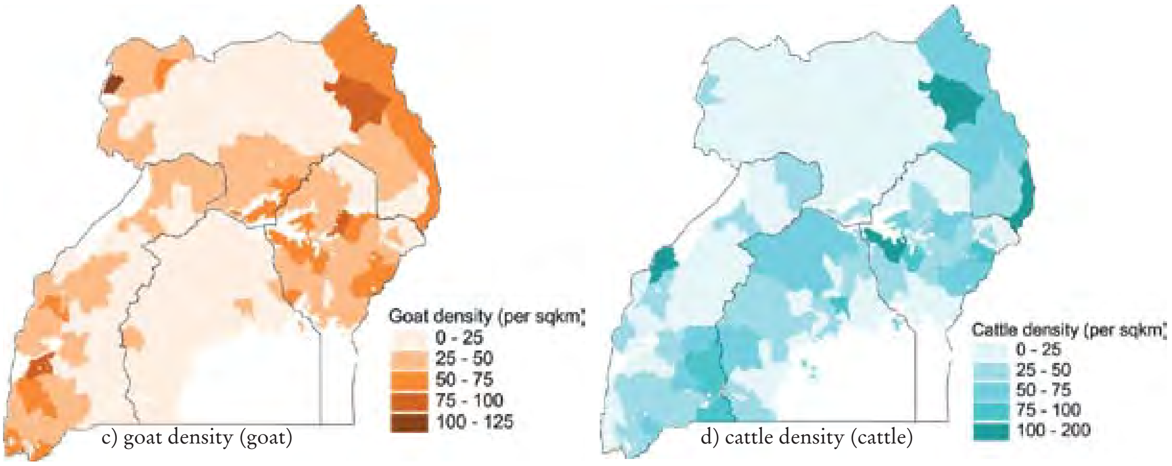
In conclusion, these variables show some degree of correlation with per-adult equivalent expenditure, have little collinearity, and, in general have the expected sign.

Figure 8. Independent variables used in the regression models.



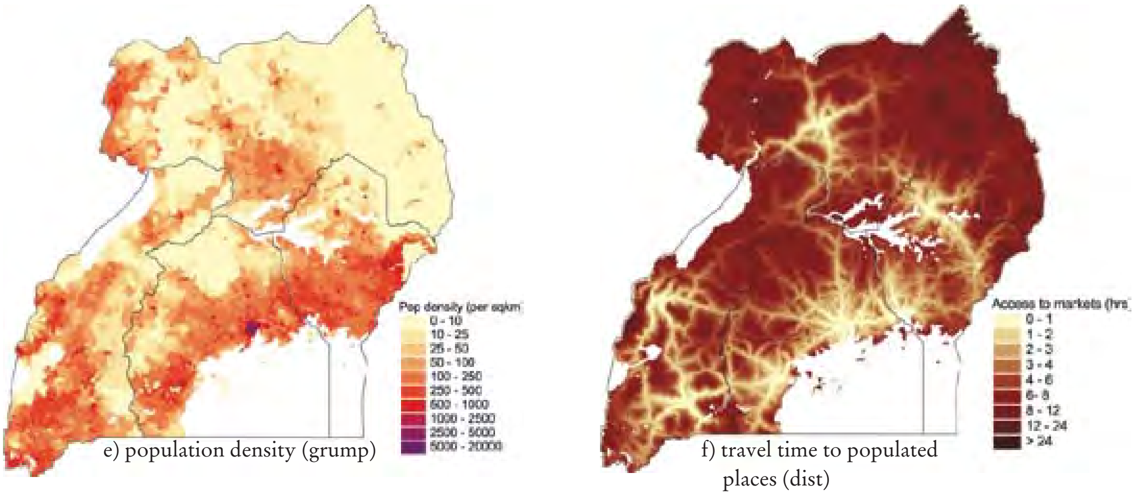
a) NDVI (ndvi)

b) Vapour Pressure Deficit (vpd)



c) goat density (goat)

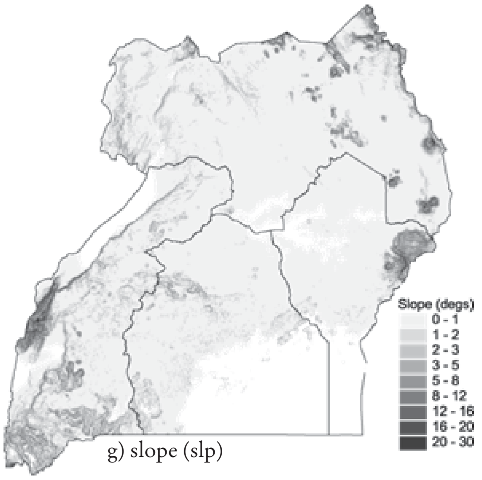
d) cattle density (cattle)



e) population density (grump)

f) travel time to populated places (dist)

Figure 8 (cont).



g) slope (slp)

Table 3. Correlation matrix for the dependent (ybct) and independent variables.

	ndvi	vpd	goat	cattle	slp	grump	dist
ybct	0.307	-0.297	-0.178	0.045	0.041	0.194	-0.241
ndvi		-0.382	-0.254	-0.143	-0.017	0.072	-0.208
vpd			0.113	-0.092	-0.509	-0.322	0.094
goat				0.273	0.096	0.13	0.176
cattle					0.031	0.084	-0.037
Slp						0.265	0.038
grump							-0.319

Table 4. Descriptive statistics for the dependent (ybct) and independent variables at 0.01 degrees resolution.

Momental statistics (n=2 088)					
Variable	**Mean**	**Std Err Mean**	**Std Dev**	**Skewness**	**Kurtosis**
ybct	5.19	0.003	0.12	-0.01	0.34
ndvi	0.52	0.001	0.07	-2.23	18.57
vpd	2.62	0.015	0.68	0.08	-0.44
goat	32.43	0.427	19.50	0.95	1.90
cattle	33.37	0.471	21.54	1.43	2.47
slp	1.13	0.032	1.48	3.02	10.92
grump	190.62	4.415	201.72	4.47	37.87
dist	254.12	3.276	149.71	1.23	3.65

OLS RESULTS

In this section the full set of regression results and diagnostics are presented for the 0.01 degree resolution analysis, and summaries for the analyses conducted at coarser resolutions.

The results show each independent variable to be significant at the 1 percent level or better (Table 5). The overall model is significant, explaining 18.52 percent (the multiple r-squared value in Table 5, expressed as a percentage) of the variability in the rural monthly adult equivalent expenditure at 0.01 degrees resolution. Throughout, R^2 rather than adjusted R^2 has been used as it is a direct measure of the model's ability to explain the variance in the data (adjusted R^2 values remove the effect of collinearity of the predictors, which was slight here).

Table 5. Descriptive statistics for the dependent (ybct) and independent variables at 0.01 degrees resolution (c. 1.1 km at the equator).

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Signifi.[1]
(Intercept)	5.155e+00	3.246e-02	158 829 .	< 2e-16	***
ndvi	3.085e-01	4.084e-02	7.555	6.23e-14	***
vpd	-3.521e-02	4.737e-03	-7.432	1.55e-13	***
goat	-6.884e-04	1.338e-04	-5.143	2.95e-07	***
cattle	3.883e-04	1.160e-04	3.347	0.000832	***
slp	-5.789e-03	1.947e-03	-2.974	0.002974	**
grump	6.071e-05	1.337e-05	4.539	5.97e-06	***
dist	-1.003e-04	1.733e-05	-5.789	8.16e-09	***

Residual standard error: 0.1067 on 2 080 degrees of freedom; multiple R-squared: 0.1852; adjusted R-squared: 0.1825; F-statistic: 67.55 on 7 and 2 080 DF; p-value: < 2.2e-16; AICc: -3 410.745.

[1] *** p<0.001 ; ** p<0.01 ; * p<0.05

The signs of the coefficients were as expected: NDVI, cattle density and population density all had a positive influence on rural monthly adult equivalent expenditure, while VPD, goat density, slope and travel time to markets all had a negative influence.

A set of standard diagnostic plots is shown in Figure 9. The graphs in Figures 9a and 9b plot the residuals against the fitted or predicted values. Ideally the variance in the residuals should be constant regardless of the predicted value and no obvious pattern in the scatter should be evident. Evidence of non-constant variance (heteroskedasticity) would appear as a fan-shaped pattern of increasing variance. Figure 9c is a q-q plot confirming that the residuals are normally distributed. The final graph, Figure 9d, measures the 'leverage', which refers to the degree to which some points unduly influence the regression model – high leverage means high influence.

Two tests were performed to see whether key assumptions of the OLS model were being violated. A formal chi-square test for non constant variance gave a value of 4.279254 (df = 1), with $p = 0.03858027$; any value of $p < 0.05$ confirms the unchanging nature of the residuals (Figures 9a and 9b). Multicollinearity in the independent variables was further tested for, using Variance Inflation Factors (VIF), which are indices that measure how much the variance of a coefficient increases because of collinearity. The smaller these VIF scores the better, with values of less than 2.0 being acceptable (Table 6). Both tests therefore confirmed that the OLS model did not violate the assumptions of OLS regression.

Figure 9. Diagnostic plots of the 0.01 degree OLS regression.

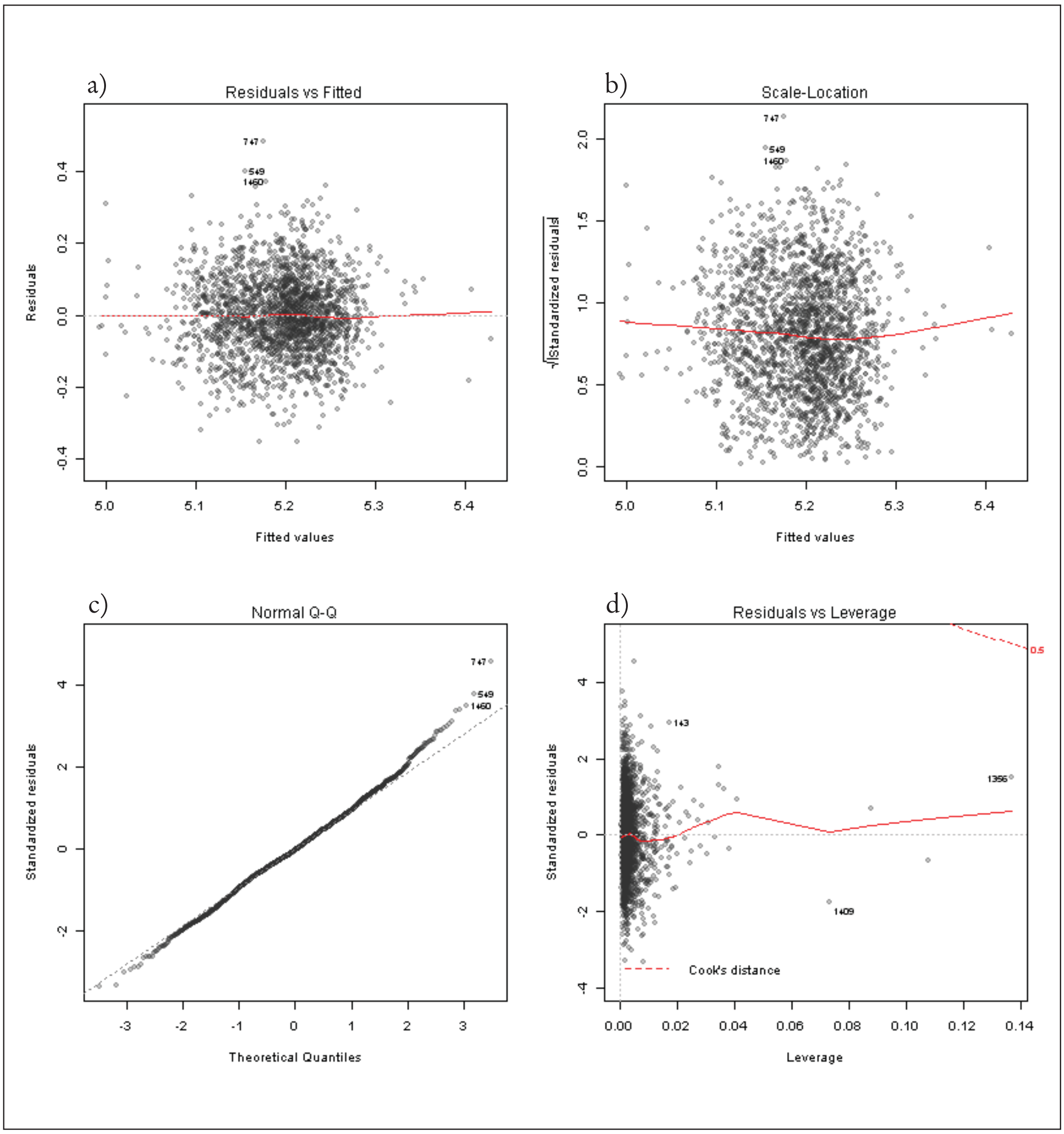


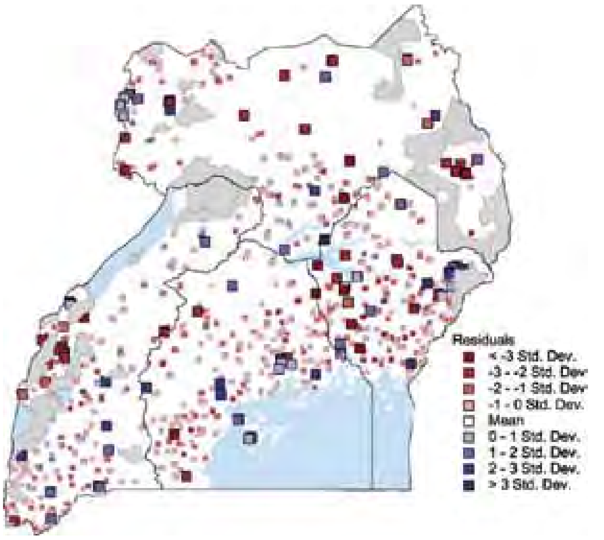
Table 6. Results of a Variance Inflation Factor (VIF) test for multicollinearity.

Variable	ndvi	vpd	goat	cattle	slp	grump	dist
VIF	1.174	1.374	1.118	1.070	1.230	1.156	1.111

Note: VIF values less than 2 indicate collinearity not to be a problem.

Figure 10 maps the standardized residuals and highlights the 110 points with high leverage as defined by the conventional cut-off of a Cook's distance greater than 4/n, where n is the number of observations (Bollen and Jackman 1990). Since these points were not spatially clustered, but rather distributed evenly across Uganda, there was no good theoretical reason to remove them.

Figure 10. Standardized residuals for the 0.01 degree OLS regression with high leverage points highlighted with larger symbols.



The relaimpo library in R was used to conduct a Lindeman, Merenda and Gold (LMG) analysis (Linderman *et al.* 1980), that quantifies the relative importance of each predictor variable in determining the model's explanatory power . The LMG analysis "computes the sequential sums of squares from the linear model…for an overall assessment by averaging over all orderings of regressors" (Grömping 2007) resulting in a decomposition of R^2 by variable. The resulting plot is shown in Figure 11. NDVI and VPD are the two most important variables at this scale for the countrywide regression, with travel time to populated centres third. These three variables accounted for almost 80 percent of the explanatory power of the model.

Finally, the coefficients from the OLS model were used to predict average rural monthly adult equivalent expenditure at 0.01 degrees resolution across Uganda. This map (which does not necessarily represent the most appropriate model or resolution for estimating the expenditure) is shown in Figure 12.

Figure 11. Estimate of the relative importance of the independent variables for the 0.01 degree OLS regression, including 95 percent confidence limits.

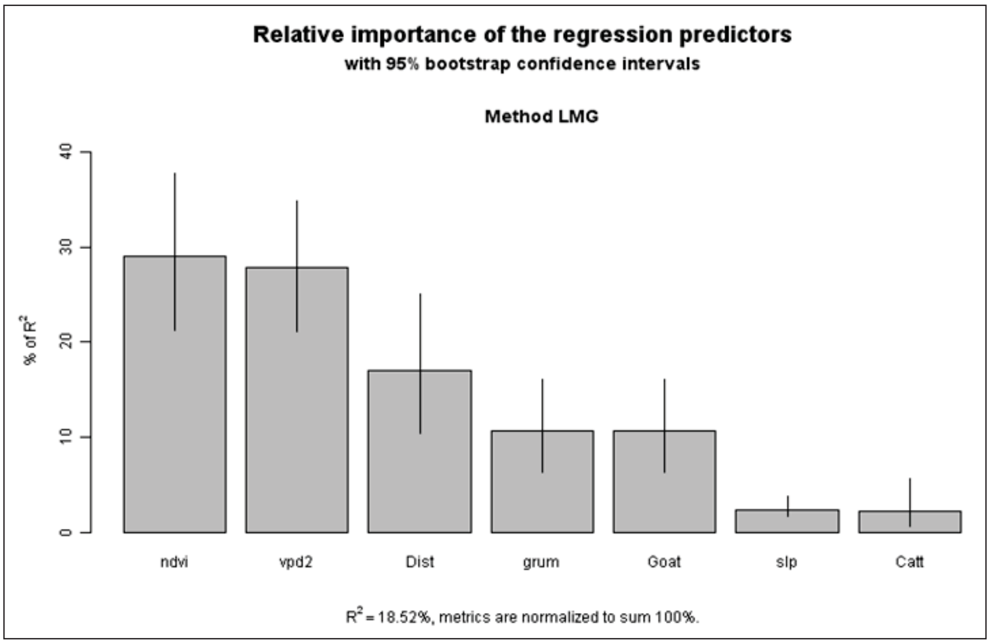
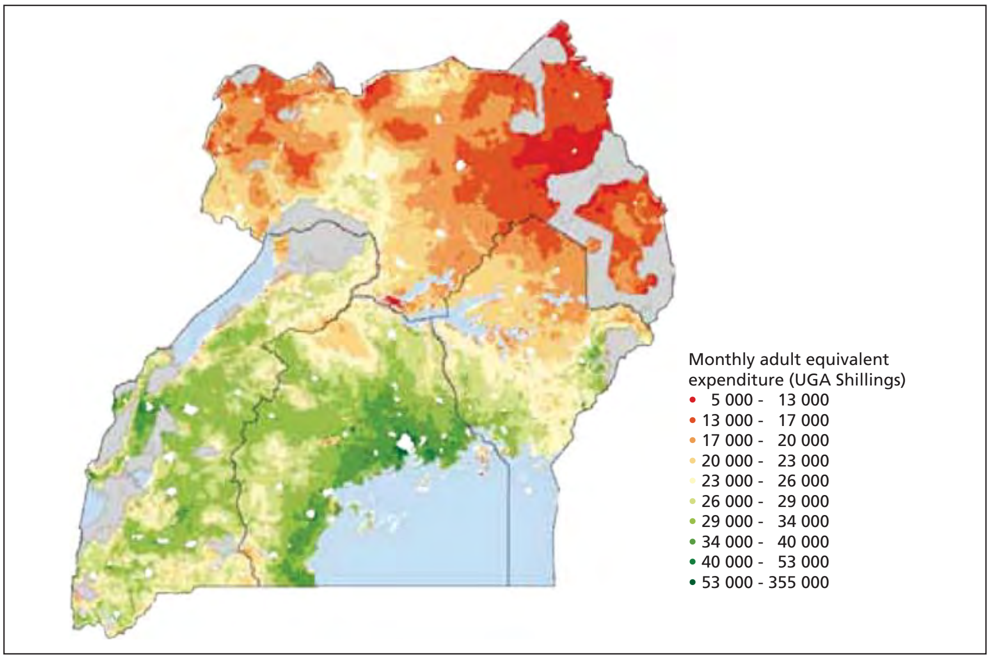


Figure 12. Predicted average rural monthly adult equivalent expenditure.



Note: based on a country-wide OLS regression model at 0.01 degrees resolution (c. 1.1 km at the equator).

The map bears a strong resemblance to the SAE map of rural monthly adult equivalent expenditure (Figure 2), with lower expenditures in the northern and especially eastern regions and higher expenditures in the southern and central regions, especially around Kampala and Lake Victoria.

The same analysis was performed at each spatial resolution, for a subset of which the OLS model outputs are given in Table 7. As a rule of thumb there should be at least 50 independent data points for each independent variable, so at least 350 data points are required in this case (7 independent variables). Any OLS results at 0.20 degrees and coarser resolutions, therefore, should be treated with caution. Focussing on the results from cell sizes of 0.01 to 0.15, Table 7 shows that R^2 tends to increase, while the sign and relative importance of the variables was more or less constant: NDVI, VPD, population density and goat density were the most important variables, while travel time to markets, slope and cattle density were generally less important.

Table 7. OLS model summary for all resolutions.

Model results			Significance[1], sign (+/-) and relative importance (1-7)						
Cell size	**Points**	**R^2**	**ndvi**	**vpd**	**goat**	**cattle**	**slp**	**grump**	**dist**
0.01	2 088	0.185	*** + **1**	*** - **2**	*** - 5	*** + 7	*** - 6	*** + 4	*** - 3
0.02	1 279	0.176	*** + 4	*** - **1**	*** - **2**	** + 6	*** - 5	*** + 3	*** - 5
0.03	1 086	0.206	*** + 3	*** - **1**	*** - **2**	** + 7	*** - 6	*** + 4	*** - 5
0.05	813	0.227	*** + 4	*** - **1**	*** - 3	** + 7	*** - 5	*** + **2**	- 6
0.10	539	0.292	*** + 4	*** - **1**	*** - 3	** + 7	*** - 6	*** + **2**	- 5
0.15	399	0.290	* + 4	*** - **1**	*** - 3	* + 7	*** - 5	*** + **2**	- 6
0.20	*280*	*0.364*	- *6*	*** - ***1***	*** - *3*	+ *7*	*** - *4*	*** + ***2***	- *5*
0.25	*206*	*0.371*	- *6*	*** - ***1***	** - *3*	+ *7*	*** - *4*	*** + ***2***	- *5*
0.30	*167*	*0.421*	- *7*	*** - ***1***	*** - *3*	+ *5*	*** - *4*	*** + ***2***	- *6*
0.35	*120*	*0.513*	- *6*	*** - ***1***	*** - *3*	** + *7*	*** - *4*	*** + ***2***	- *5*
0.40	*103*	*0.409*	+ *4*	*** - ***1***	** - *3*	+ *7*	* - *6*	** + ***2***	+ *5*
0.45	*82*	*0.587*	* - *7*	*** - ***1***	** - *3*	*** + *6*	*** - ***2***	** + *4*	- *5*
0.50	*75*	*0.527*	+ *6*	*** - ***1***	- *3*	+ *7*	* - *4*	* + ***2***	- *5*
0.75	*36*	*0.614*	- *5*	*** - ***1***	- *4*	+ *7*	*** - *3*	+ ***2***	- *6*

[1] *** $p<0.001$; ** $p<0.01$; * $p<0.05$

Note: Most important variables in bold. Table rows in italics are for regressions with fewer than the minimum recommended number of data points (see text).

REGIONAL OLS RESULTS

Analyses of the six regional sub-sets of the data were carried out in the same way as described above for the single country-wide regression. Not all results are presented here since the regional models encountered the same problem of insufficient data points as did the country-wide model at coarser spatial resolutions. The main purpose here was to determine whether there were differences in the sign and significance of the regression coefficients, and of the relative importance of the regression variables, in the different zonations used. Table 8 summarises these three features for the six regions at 0.01, 0.05 and 0.10 degrees spatial resolution.

It follows from the description earlier of the zonation schemes used here that there is considerable overlap between the ones containing large numbers of data points; in particular the humid and sub-humid climate zone; the mixed crop and livestock systems; and the intersection of these - the dominant mixed, humid and sub-humid system (Figure 4). It is therefore perhaps not surprising that the regional OLS results are similar to each other for these zones, and in some respects to the OLS country-wide regression. The signs of the coefficients were consistent across the regions and across resolutions; VPD and NDVI were the two most important variables while cattle and slope were consistently the least important. VPD was nearly always the most important variable in the OLS country-wide model, but the 2nd and 3rd placed variables varied by region and resolution.

Three of the regional analyses (the climate zones: arid and semi-arid, and temperate/tropical highland; and the farming system: livestock-only) lacked data points at all three resolutions.

As in the OLS results, the R^2 values generally increased as the cell size increased and the number of data points decreased. This perceived improvement in the model at larger cell sizes needs to be treated with caution but may be important. Random data showing no relationship between environmental variables and household expenditure would not show any improvement in r-squared values if they were progressively combined by averaging, as here, whereas any real relationship between these two variables is likely to be quite noisy at the finest spatial resolution and less noisy at aggregated resolutions, the effect of aggregation being to cancel out noise, and hence to reveal more of the true ‘signal’

The predicted expenditure for all six regions was computed and then combined to create three maps of expenditure at 0.01 degrees resolution as follows: (i) climate – a combination of models for the arid and semi-arid, temperate and tropical highlands and humid and sub-humid regions; (ii) farming – a combination of models from the livestock-only and mixed systems; and (iii) dominant – the mixed, humid and sub-humid region. These three maps are shown in Figure 13. As before, it is not implied that any of these represents the most appropriate model or resolution for estimating household expenditure.

The ‘climate zones’ and ‘farming systems’ regions cover the whole country (except for the ‘urban’ and ‘other’ farming systems), while the ‘dominant’ system map covers only the central area of Uganda. All three maps are similar to the SAE map of expenditure and to the OLS country-wide predictions. The ‘climate zones’ models capture more extreme ranges of expenditure than do the others, i.e. the northeast area of very low expenditure and the southwest area of very high expenditure. This is similar to the expenditure pattern seen in the SAE map (Figure 2).

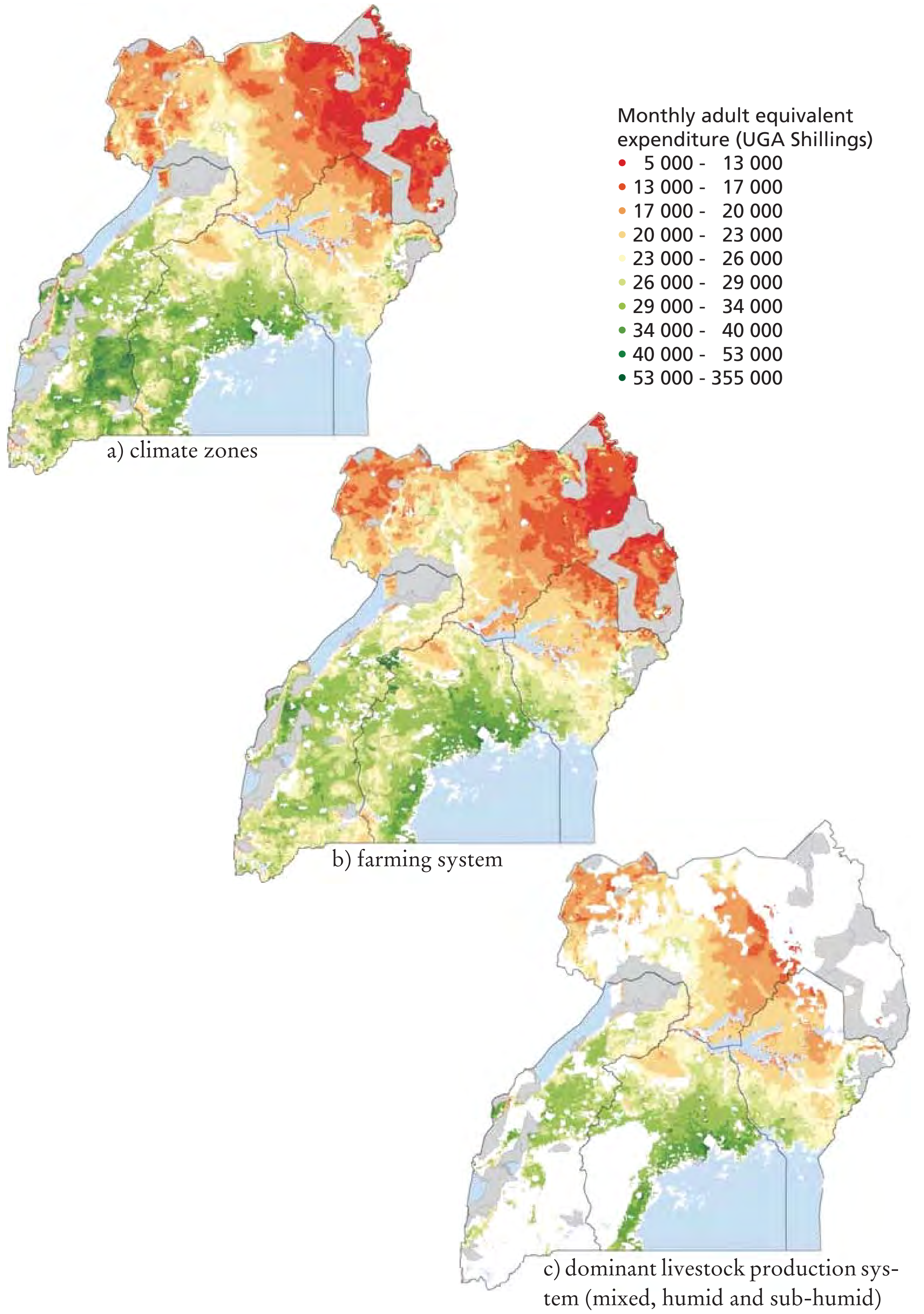
Table 8. Summaries for regional models at selected resolutions.

Model name and results			Significance[§], sign (+/-) and relative importance (1-7)							
Model	**Cell size**	**Points**	**R^2**	**ndvi**	**vpd**	**goat**	**cattle**	**slp**	**grump**	**Dist**
C1 Arid system	0.01	296	0.310	*** **+** **1**	- 3	- 5	*** + 4	+ 7	- 6	** **-** **2**
	0.05	110	0.294	+ 3	**-** **2**	* 5	+ 6	+ 7	+ 4	** **-** **1**
	0.10	87	0.362	** **+** **2**	- 3	** - 7	+ 5	- 6	+ 4	**-** **1**
C2 Temperate system	0.01	1 404	0.174	*** **+** **2**	*** **-** **1**	*** - 5	- 7	*** - 6	*** + 4	*** - 3
	0.05	546	0.235	** **+** **2**	*** **-** **1**	** - 4	+ 7	*** - 5	*** + 3	*** - 6
	0.10	352	0.335	*** **+** **2**	*** **-** **1**	** - 4	+ 7	* - 6	*** + 3	- 5
C3 Humid & sub-humid system	0.01	291	0.183	+ 6	- 7	*** **-** **2**	+ 4	- 5	+ 3	*** **-** **1**
	0.05	99	0.322	**+** **2**	* + 4	- 3	- 7	** + 5	- 6	*** **-** **1**
	0.10	68	0.340	+ 3	- 7	**-** **2**	- 5	+ 6	- 4	*** **-** **1**
F1 Livestock-only system	0.01	73	0.368	*** **+** **1**	**-** **2**	- 6	** + 3	+ 7	+ 4	- 5
	0.05	*25*	*0.390*	+ *4*	**-** ***2***	- *5*	**+** ***1***	+ *3*	+ *6*	+ *7*
	0.10	*19*	*0.357*	+ *nd*	- *nd*	- *nd*	+ *nd*	- *nd*	+ *nd*	+ *nd*
F2 Mixed system	0.01	1 918	0.176	*** **+** **2**	*** **-** **1**	*** - 4	** + 7	** - 6	*** + 5	*** - 3
	0.05	730	0.216	*** **+** **2**	*** **-** **1**	*** - 3	* + 7	*** - 6	*** + 5	** - 4
	0.10	452	0.318	**+** **1**	**-** **2**	- 4	+ 7	- 6	+ 5	- 3
Mixed, humid & sub humid system	0.01	1 357	0.168	*** **+** **2**	*** **-** **1**	*** - 4	+ 7	*** - 6	*** + 5	*** - 3
	0.05	531	0.231	** **+** **2**	*** **-** **1**	*** - 3	+ 7	*** - 5	*** + 4	+ 6
	0.10	343	0.323	*** **+** **2**	*** **-** **1**	*** - 4	+ 7	- 6	*** + 3	- 5

[§] *** p<0.001 ; ** p<0.01 ; * p<0.05

Note: Most important variables in bold, table rows in italics signify possibly insufficient numbers of data points.

Figure 13. Predicted average rural monthly adult equivalent expenditure based on regional models at 0.01 degrees resolution (c. 1.1 km at the equator).



a) climate zones

b) farming system

c) dominant livestock production system (mixed, humid and sub-humid)

Note: Areas of no prediction are in white.

GWR RESULTS

The first stage of GWR calibrates the model using a cross validation approach to determine the best bandwidth or kernel size. The GWR model is then run at that bandwidth, producing the full range of outputs and tests of significance. The model is next applied to all rural pixels in Uganda to predict average rural monthly adult equivalent expenditure for each pixel, as before. Finally the coefficients and their significance levels are mapped and interpolated.

The results at 0.01 degree resolution are shown in Table 9. The first stage results indicated that the optimal kernel size should include 807 (38.7 percent) of the 2 088 data points available to develop a single regression model for each point.

Table 9 provides the quartiles of the distribution of each coefficient as it varied across the dataset, and also gives the OLS coefficient (named 'Global') for comparison. The results show that the GWR results do vary across the region with all coefficients (except the intercept) ranging from negative to positive.

The regression outputs, given as footnotes to Table 9, can be compared with the OLS results presented in Table 5 (though see notes of caution below, regarding the use of these internal statistics for comparing different models). The GWR model has a lower sigma, lower AICc and higher R^2 value than the country-wide OLS model. An ANOVA rejected the null hypothesis that the GWR model offered no improvement over the OLS model (F = 3.9959, df1 = 767.219, df2 = 2 046.027, p-value <2.2e-16).

The GWR model outputs make comparisons of the sign, significance and importance of variables between these and OLS models difficult, but the GWR procedure in R produces a series of statistics designed to compare the GWR model with the OLS model (Table 10): R^2 (the higher the better), AICc (the lower the better) and significance levels based on the p values from two different F-tests, F1 (Leung *et al.* 2000) and F2 (Fotheringham *et al.* 2002) that compare the GWR against the OLS model. The results at coarser resolutions are not presented because there are insufficient data points to describe the GWR kernels for them.

Table 9. Summary of GWR coefficient estimates at 0.01 degree resolution (c. 1.1 km at the equator), based on a kernel size of 807 data points (38.7 percent of the 2 088 data points available).

Variable	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
(intercept)	4.85e+00	5.07e+00	5.24e+00	5.35e+00	5.61e+00	5.1555
ndvi	-4.04e-01	6.35e-02	1.63e-01	2.76e-01	6.11e-01	0.3085
vpd	-9.96e-02	-5.79e-02	-4.27e-02	-1.31e-02	4.85e-02	-0.0352
goat	-1.74e-03	-5.40e-04	-2.13e-04	3.75e-04	7.62e-04	-0.0007
cattle	-8.87e-04	-9.48e-05	3.34e-04	6.89e-04	1.07e-03	0.0004
slp	-2.69e-02	-1.64e-02	-7.17e-03	-4.67e-03	1.00e-02	-0.0058
grump	-9.46e-07	2.31e-05	4.59e-05	9.46e-05	1.78e-04	0.0001
dist	-3.05e-04	-1.55e-04	-7.83e-05	-5.02e-05	1.29e-05	-0.0001

Effective number of parameters: 55 97674; effective degrees of freedom: 2 032.023; sigma: 0.1031564; AICc: 3 524.839; residual sum of squares: 21 62324; GWR multiple R^2: 0.2555 (compared with the OLS multiple R^2 of 0.1852).

Both the R^2 and AICc scores suggest that GWR out-performs OLS across all resolutions, even when accounting for the added model complexity in GWR. However, statistics like these, estimated within a model, should really only be used in similar models, for example, the AICc is generally used to see if a particular predictor variable improves the fit of the model. Comparing R^2 and AICc values across models with different structures, different sets of variables and, most importantly, different numbers of data points is thus problematic. Consequently, in the following section, more robust comparisons are made between the different models.

GWR suffers from a lack of data points at resolutions above 0.05 degrees resolution. Figure 14 shows the predicted expenditure using the GWR model at 0.01 degrees resolution. Again, the resulting map is very similar to the results of the previous models, and to the SAE map.

Having run all models at all resolutions it was then possible to perform a direct comparison of the predictions across all models to identify the best performing model and resolution.

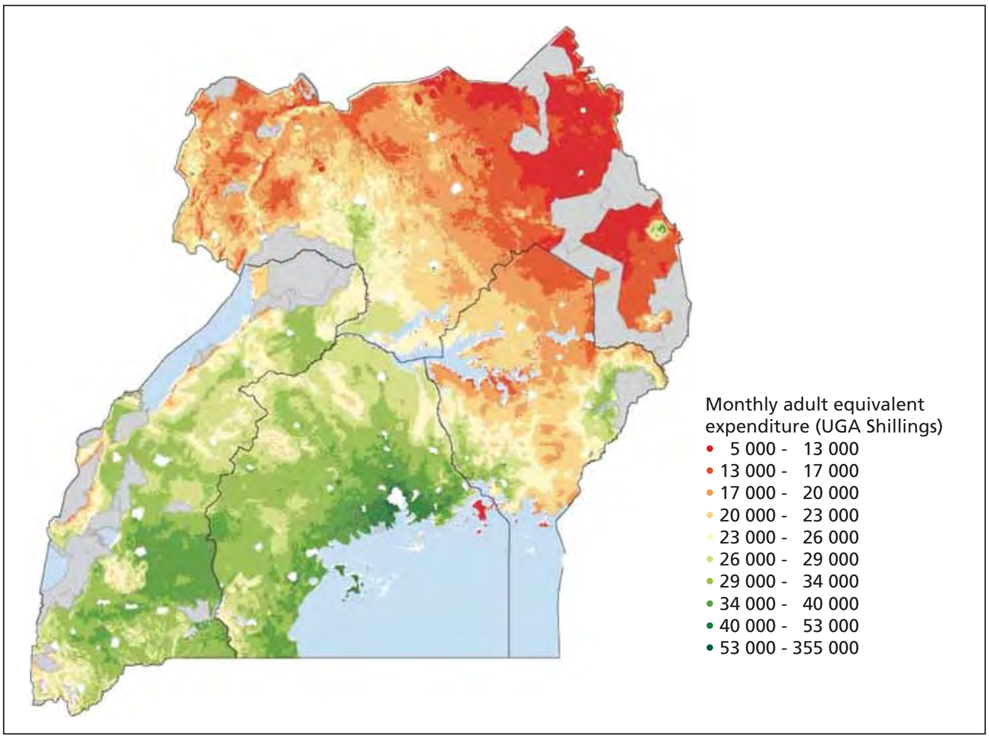
Table 10. GWR and OLS model comparison.

Model scale		Kernel size		R^2		AICc		Significant improvement?[1]	
Cell size	**Points**	**Points**	**As a %**	**GWR**	**OLS**	**GWR**	**OLS**	**F1**	**F2**
0.01	2 088	807	39%	0.256	0.185	-3 525	-3 411	***	***
0.02	1 279	419	33%	0.300	0.176	-363	-248	***	***
0.03	1 086	*201*	*19%*	*0.376*	0.206	*672*	762	***	***
0.05	813	387	48%	0.351	0.227	1 408	1 494	***	***
0.10	539	*178*	*33%*	*0.472*	0.292	*1 532*	1 599	***	***
0.15	399	*172*	*43%*	*0.460*	0.290	*1 962*	2 009	***	***

[1] *** $p<0.001$; ** $p<0.01$; * $p<0.05$

Note AICc scores should be compared at the same resolution, not across resolutions. GWR data in italics are based on few data points

Figure 14. Predicted average rural monthly adult equivalent expenditure based on the GWR model at 0.01 degrees resolution.



GOODNESS OF FIT METRICS FOR ALL REGRESSION MODELS AND THE SMALL AREAS ESTIMATES

Often R^2 or adjusted-R^2 values generated within regression models are used to compare models. However, when comparing regression models in which the dependent variable has been transformed in different ways, which used different sets of data points, and which include different combinations of independent variables then the model R^2 is not a reliable guide in comparing model quality. In such cases direct comparisons between the predicted values and the observations should be used, such at the R^2 estimate for the relationship between observed and model-predicted values, RMSE and other, related metrics.

Although the residual standard error (or Sigma) from a regression model is effectively the same as the RMSE, Sigmas cannot be compared directly across the models produced here because each model is based on a different transformation of the dependent variable. So, instead, after back transforming the predicted rural monthly adult equivalent expenditure for each of the *n* pixels containing rural households, the RMSE in Ugandan Shillings was estimated for each model at each resolution as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}}$$

The mean absolute error (MAE) and mean absolute percentage error (MAPE) are two other measures of fit that are less sensitive to outliers than is the RMSE:

$$MAE = \frac{\sum_{i=1}^{n} \left| predicted_i - actual_i \right|}{n}$$

$$MAPE = \left(\frac{\sum_{i=1}^{n} \left| predicted_i - actual_i \right|}{actual_i} \right) . \frac{1}{n}$$

Finally, for completeness, the R^2 value was computed from the plot of observed vs. expected expenditure for all data points, at all resolutions. However, this suffers from the same sensitivity to outliers as does the RMSE.

At each resolution the country-wide OLS, regional OLS and GWR models were bootstrapped. Each regression model was run 1 000 times with bootstrapped samples from the original dataset to obtain a distribution of the four metrics, which were then used to generate unbiased estimates and standard errors, shown in Table 11.

The same four metrics were estimated for the SAE expenditure maps at district, county and sub-county levels (Table 12). These could only be computed based on the administrative units that contained rural household points, just as the regression model used only those pixels that contained household points. The average administrative unit size (with standard errors) was estimated for each SAE, and two extreme outliers were removed from the sub-county level and two from the county level SAE results before computing the metrics[6].

Figure 15a shows the results for MAE, and Figure 15b, shows the same results in greater detail for the finer resolutions (up to 300 km^2, beyond which there were insufficient records to allow robust predictions to be made – see Table 11). Model performance is plotted against average pixel size in square kilometres, demonstrating the trade-off between model accuracy and the spatial resolution. The SAE results are also included on the graph, although it was not possible to compute standard errors for these (though standard errors around the average administrative unit area are given). In all cases the results for the regional OLS models lay between the country wide OLS and GWR results, though for clarity these have been omitted from Figure 15.

The results show that the GWR predictions were better than the regional OLS models, which, in turn, were better than the country-wide OLS. They also show that the country-wide OLS and GWR models have similar metric scores to the SAE models at cell sizes that were comparable to the district and country scales. However the OLS and in particular the GWR models had significantly better metric scores than the sub-county SAE models at comparable scales. For example, the sub-county RMSE was 16 614, comparable to the 0.02 degrees resolution GWR model, with an RMSE equal to 16 339; a 44-fold increase in spatial precision. For MAE and

---

[6] Sub counties 406206 and 205103 and their corresponding counties 4062 and 2051. There are no SAE for the corresponding districts 406 and 205.

MAPE the comparable GWR resolutions were 0.03 and 0.05 degrees; a 20- or 7-fold increase respectively, and for R^2 it was 0.01 degrees (a 178-fold increase)

Considering all the metrics in Table 11 and the shape of the curve in Figure 15b, a cell size of 0.05 degrees, covering approximately 31 km^2, or 5.5 × 5.5 km, results in a conservative trade-off between spatial precision and the predictive accuracy of the model. At this resolution (as with almost all others), GWR gives the best result followed by the regional OLS models for the dominant (mixed, humid and sub-humid) livestock production system, and finally the country-wide OLS model. Figure 16 shows the predicted average monthly rural household expenditure for the GWR model at 0.05 degrees resolution. These estimates have lower or comparable errors to the finest SAE map and are over seven times as detailed as the SAE rural monthly adult equivalent expenditure estimates at sub-county level. The summary results for the 0.05 degree GWR model are also shown.

Table 11. Goodness of fit metrics for GWR and OLS models at each resolution.

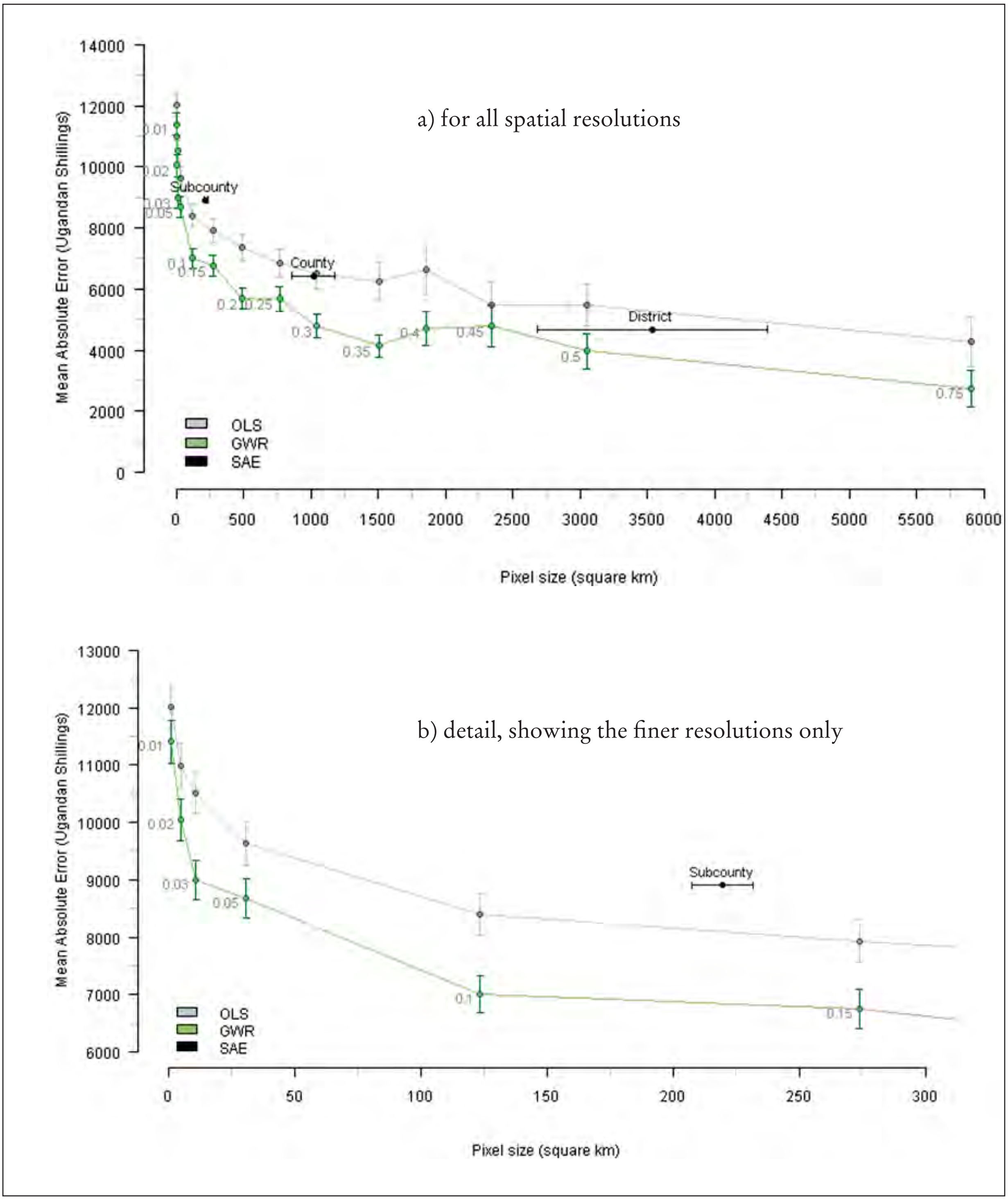
			RMSE (UGA Shillings)		MAE (UGA Shillings)		MAPE (%)		R^2 (for Obs vs Exp)	
Cell size (degrees)	**Records**	**Sq km**	**GWR**	**OLS**	**GWR**	**OLS**	**GWR**	**OLS**	**GWR**	**OLS**
0.01	2 088	1.2	20 462±1 563	21 034±1 531	11 408±371	12 024±382	37.4±0.7	40.2±0.8	0.17±0.02	0.11±0.02
0.02	1 279	4.9	16 339±953	17 563±968	10 044±368	10 991±387	33.5±0.9	37.7±1.0	0.25±0.03	0.13±0.02
0.03	1 086	11.1	14 053±784	16 091±824	8 996±333	10 518±355	30.4±0.9	36.5±1.0	0.37±0.03	0.17±0.03
0.05	813	30.9	12 893±713	14 173±719	8 680±348	9 637±374	30.6±1.0	34.1±1.1	0.34±0.04	0.22±0.04
0.10	539	124	9 866±602	11 772±644	7 001±316	8 394±366	23.8±0.9	29.3±1.2	0.51±0.04	0.30±0.04
0.15	399	274	9 170±455	10 854±520	6 746±340	7 933±378	24.1±1.4	29.0±1.5	0.51±0.04	0.32±0.05
0.20	*280*	*493*	*7 690±433*	*9 840±572*	*5 700±354*	*7 347±426*	*20.5±1.3*	*27.3±1.7*	*0.64±0.05*	*0.40±0.07*
0.25	*206*	*770*	*7 660±498*	*9 236±576*	*5 680±410*	*6 846±477*	*21.9±2.0*	*27.0±2.6*	*0.61±0.06*	*0.43±0.08*
0.30	*167*	*1 047*	*6 492±519*	*8 469±577*	*4 799±384*	*6 490±496*	*17.1±1.5*	*23.7±2.1*	*0.69±0.04*	*0.46±0.06*
0.35	*120*	*1 504*	*5 484±424*	*8 646±976*	*4 132±368*	*6 266±630*	*14.7±1.6*	*22.9±2.6*	*0.83±0.04*	*0.58±0.08*
0.40	*103*	*1 854*	*6 500±747*	*9 108±1 166*	*4 727±555*	*6 622±782*	*17.6±2.4*	*25.7±3.3*	*0.72±0.05*	*0.46±0.08*
0.45	*82*	*2 342*	*6 731±1 080*	*7 625±1 243*	*4 784±678*	*5 500±734*	*16.4±2.5*	*18.6±2.7*	*0.76±0.08*	*0.69±0.09*
0.50	*75*	*3 051*	*5 561±884*	*7 478±976*	*3 960±579*	*5 499±687*	*16.2±2.6*	*23.1±3.5*	*0.78±0.08*	*0.61±0.10*
0.75	*36*	*5 903*	*3 743±730*	*5 386±903*	*2 753±599*	*4 283±830*	*12.6±3.6*	*19.3±4.8*	*0.87±0.05*	*0.72±0.08*

Note: Rows in italics are models with few data points.

Table 12. Goodness of fit metrics for the Small Area Estimates at each scale.

SAE unit	Records	Sq km	RMSE	MAE	MAPE (%)	R^2
Sub County	528	220±12	16 614	8 910	29.9	0.14
County	144	1 018±159	9 109	6 432	20.7	0.49
District	53	3 537±859	6 153	4 669	17.4	0.68

Figure 15. Mean Absolute Error, with bootstrapped standard errors over 1 000 replications, for country-wide GWR and OLS regression models at all resolutions, and for SAE.



Note: GWR points are labelled with the cell size in degrees. The horizontal error bars on the SAE values show the standard errors of the mean area of the administrative units.

Figure 16. Predicted average rural monthly adult equivalent expenditure based on the best performing method, a Geographically Weighted Regression (bandwidth = 387 neighbours) model at 0.05 degrees resolution (c. 5.5 km at the equator).

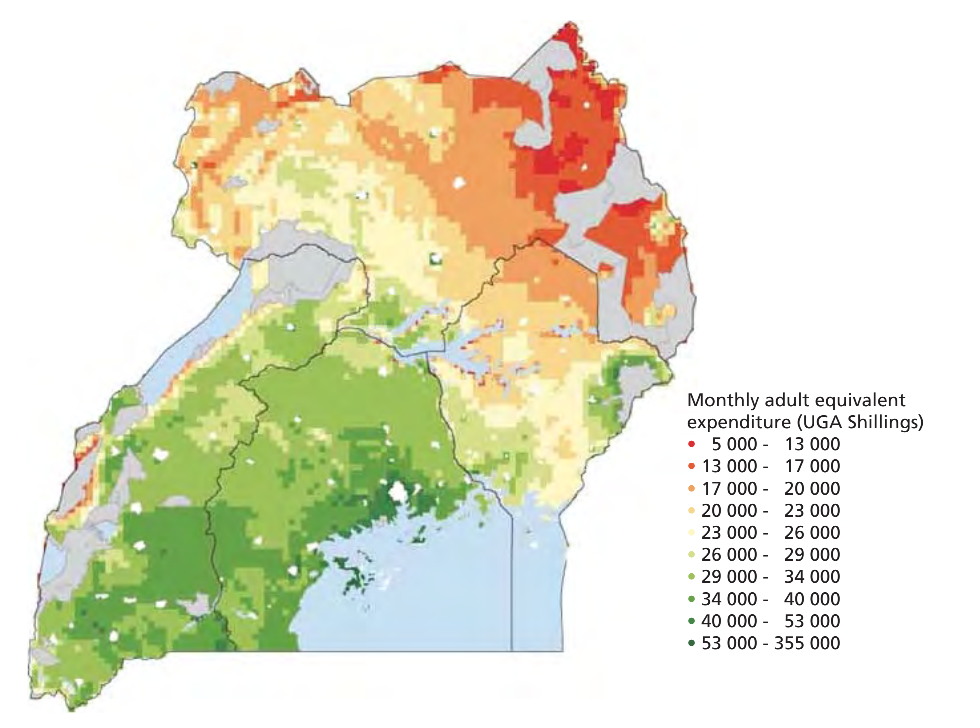


Table 13. Summary of GWR coefficient estimates at 0.05 degree resolution (c. 5.5 km at the equator), based on a kernel size of 387 data points (47.6 percent of the 813 data points available).

Variable	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
(intercept)	1.12e+01	1.21e+01	1.27e+01	1.31e+01	1.37e+01	2.8100
ndvi	-9.46e-01	2.37e-01	5.34e-01	9.74e-01	1.71e+00	0.7700
vpd	-5.29e-01	-3.51e-01	-2.23e-01	-9.03e-02	2.63e-01	-0.2994
goat	-1.15e-02	-6.16e-03	-1.23e-03	3.53e-03	8.31e-03	-0.0062
cattle	-6.14e-03	-1.21e-03	2.77e-03	4.21e-03	7.07e-03	0.0023
slp	-6.61e-01	-3.17e-01	-1.55e-01	-6.22e-02	1.77e-01	-0.1149
grump	1.21e-04	3.75e-04	5.58e-04	6.81e-04	1.90e-03	0.0006
dist	-8.04e-04	-2.28e-04	2.97e-05	1.59e-04	6.57e-04	-0.0001

Effective number of parameters: 43.80606; effective degrees of freedom: 769.194; sigma: 0.5493839; AICc: 1 367.77; residual sum of squares: 245.3818; GWR multiple R^2: 0.3514 (compared with the OLS multiple R^2 of 0.2271).

SPATIAL VARIATION IN THE GWR COEFFICIENTS

This section explores whether significant spatial variation in the GWR coefficients was present. Such variation would imply that the dependent variables relate to rural monthly adult equivalent expenditure in different ways in different areas of Uganda. In extreme cases, strong variation may indicate model misspecification (i.e. the need to use different dependent variables in a particular location). Although such variation can be investigate at a range of spatial resolutions and bandwidths, here, the analysis is presented only for the 'best' model; at 0.05 degrees resolution with a bandwidth of 387 neighbours.

Leung *et al.* (2000) developed a formal F test for GWR to determine if the variation in the GWR coefficients is significant. The results (Table 14), given below, show that NDVI, with p = 6.6 percent, fell short of significant at the usually accepted 5 percent level, but that all the other coefficients were significant at the 0.1 percent level or better. In other words there is significant spatial variation in most of the GWR coefficients.

Table 14. Test for spatial variation in the GWR coefficients based on the method.

Variable	F statistic	Numerator d.f.	Denominator d.f	Pr(>)	Significance[1]
(intercept)	2.5323	86.5142	780.28	2.182e-11	***
ndvi	1.3833	37.1461	780.28	0.06611	(*)
vpd	4.8308	279.4226	780.28	<2.2e-16	***
goat	4.7707	292.2597	780.28	<2.2e-16	***
cattle	3.6694	213.6288	780.28	<2.2e-16	***
slp	2.7149	81.4713	780.28	1.548e-12	***
grump	3.0879	24.9101	780.28	7.875e-07	***
dist	1.9658	145.0732	780.28	5.383e-09	***

[1] *** p<0.001 ; ** p<0.01 ; * p<0.05; (*) p<0.1.

Based on this the coefficients were mapped, along with their significance levels (based on a t-test). The maps in Figure 17 are arranged in pairs with the coefficient maps presented alongside the significance maps, for each parameter in turn, showing the following.

- The spatial variation in each coefficient using a red-blue (low-high) bipolar colour scheme based on standard deviations of the coefficient values.
- The zero value (where it exists) of the GWR coefficient as a green 'contour' line, to demarcate where the coefficient switches from a positive to a negative effect. Negative areas are generally in red shades, but population density is the exception where there are no negative values.
- The country-wide OLS parameter value as a black 'contour' line.
- The regions where the coefficients are significant are shaded in green; the darker the shade, the higher the level of significance.

Unlike the previous maps, protected areas, urban areas and lake overlays are not shown, as it would further complicate the maps, without aiding interpretation. The regional boundaries are given as a locational aid.

Figure 17. Mapped coefficient values (left-hand side) and their significance levels (right hand side) for the 0.05 degrees resolution GWR model.

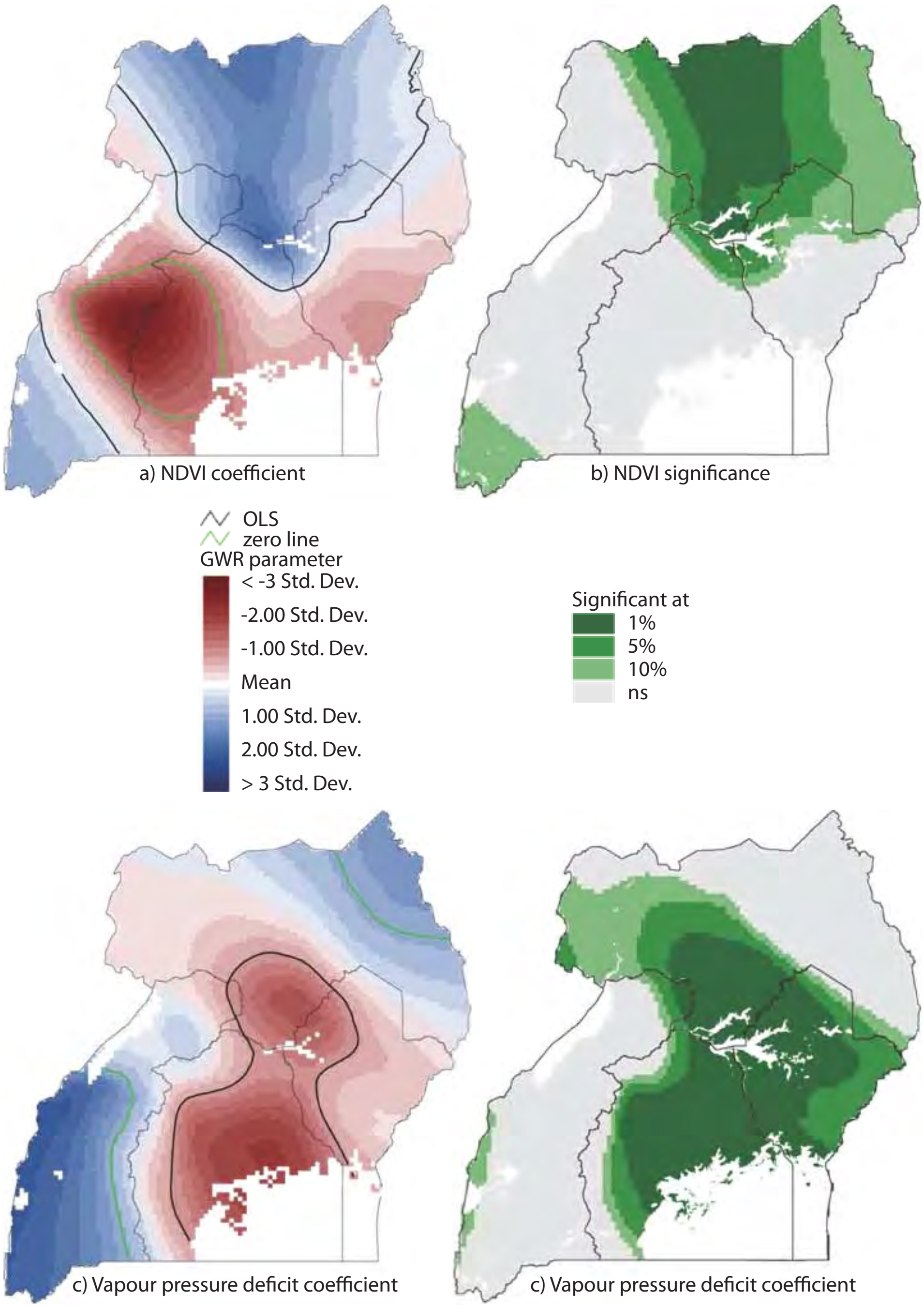


Figure 17. Continued.

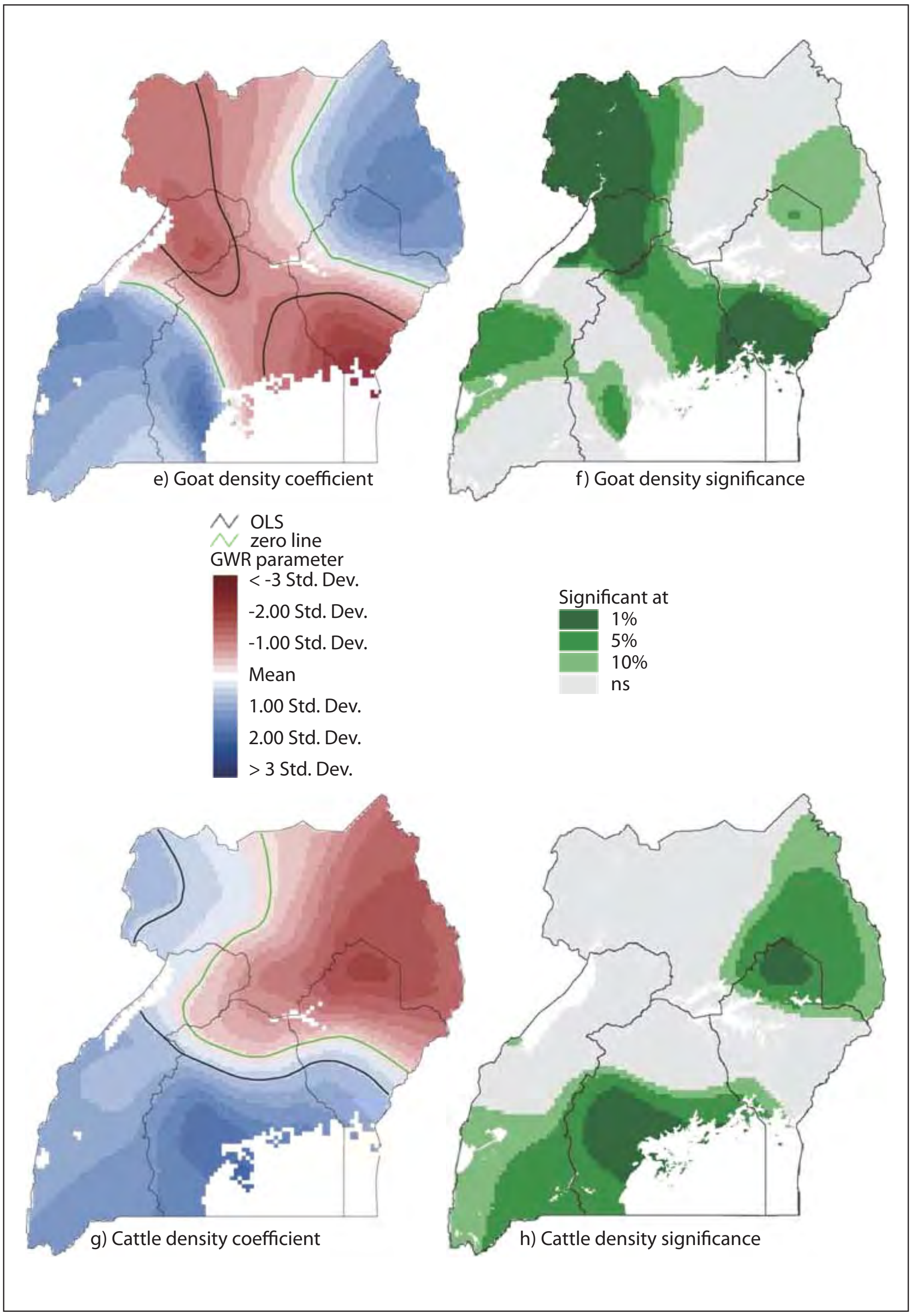
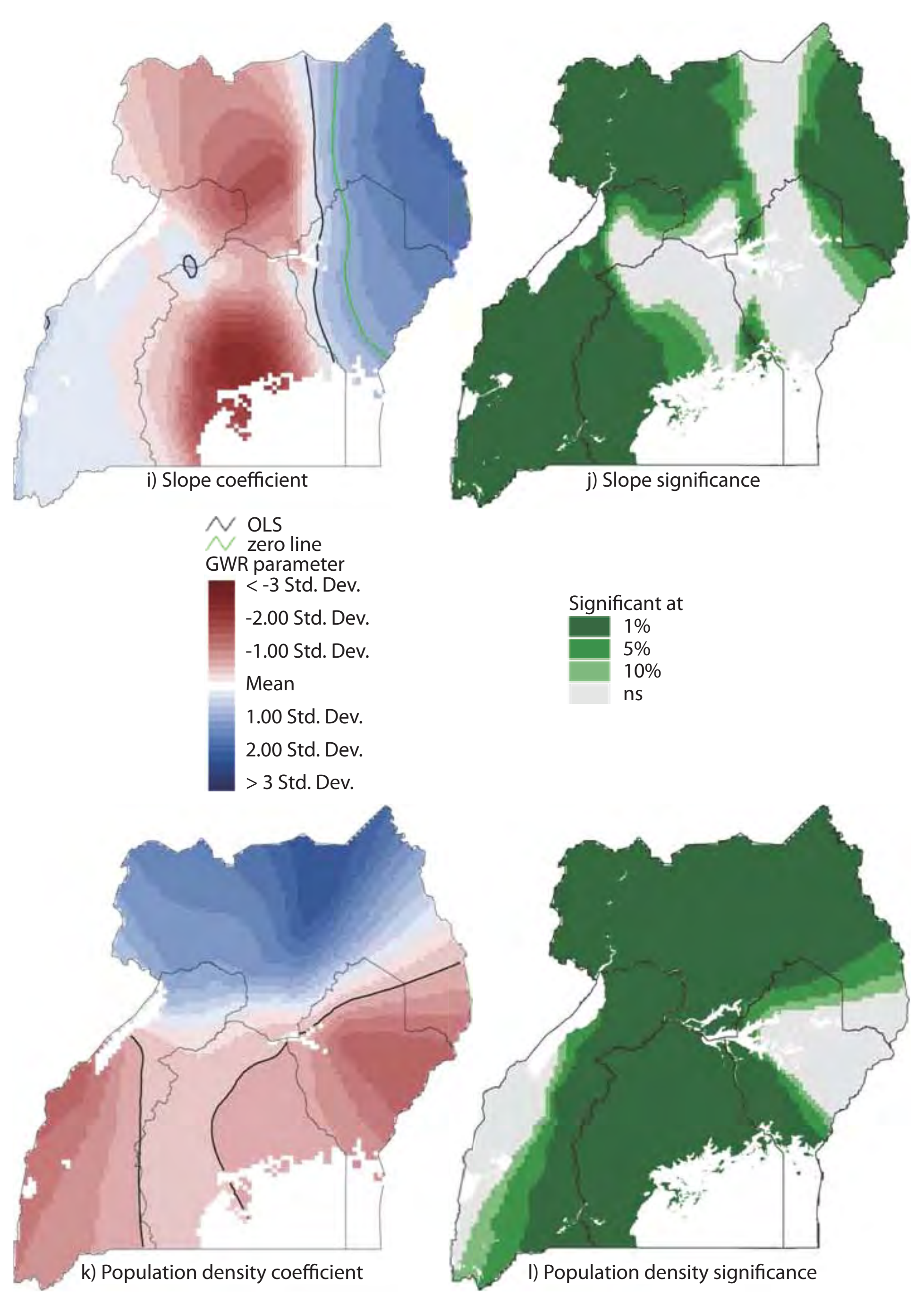


Figure 17. Continued.



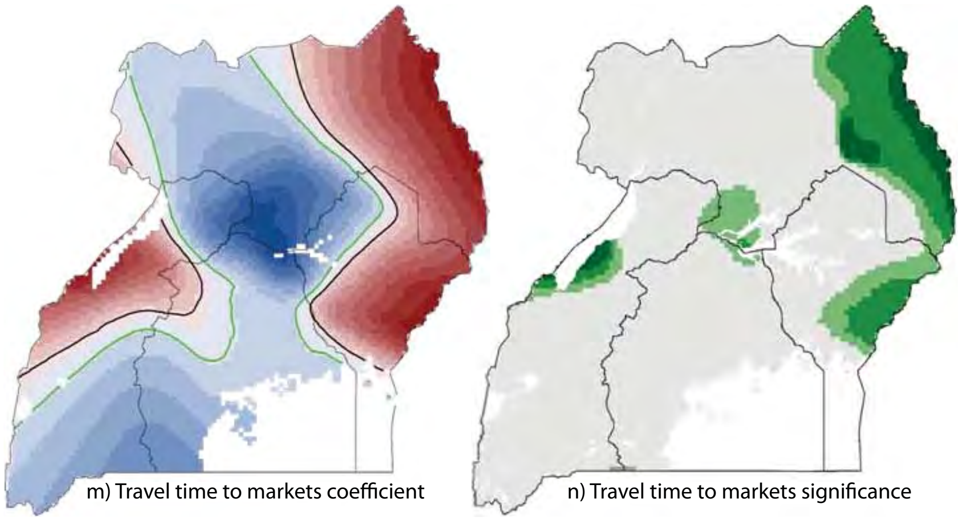
i) Slope coefficient

j) Slope significance

k) Population density coefficient

l) Population density significance

Figure 17. Continued.


m) Travel time to markets coefficient
n) Travel time to markets significance

INTERPRETING THE GWR COEFFICIENTS

In this section, possible explanations are offered for the observed patterns in each GWR parameter map, and its attendant significance estimate, in an attempt to interpret what the resulting patterns may say about how rural poverty is related to environmental conditions in different areas. Whilst we talk here about positive or negative regression coefficients we do not mean to imply the causation usually associated with regression results. The results still only confirm correlation rather than causation.

NDVI

The expected influence of NDVI would be positive, with greater vegetation vigour corresponding to higher levels of expenditure. The results (Figures 17a and 17b) showed strong positive correlations in all areas other than a patch in the centre/southwest of the country (within the green contour of Figure 17a).

Higher NDVI values broadly indicate richer vegetation growth, longer growing season(s) and higher rainfall. In the drier areas in the north, northeast and extreme southwest the coefficient of NDVI values on expenditure is positive, as one would expect, and it is in these areas that the GWR parameter is significant (according to the t-test). In the much greener areas of the central part of Uganda, the NDVI coefficient is not significant so the reversal of its influence in these areas is not of concern.

These patterns suggest that in this model there is a saturation level in terms of vegetation vigour, beyond which there is little or no benefit. Large areas of Uganda are well-served in terms of length of growing period (which is highly correlated with the annual integrated NDVI). It is only in the relatively dry areas that NDVI is likely to be limiting for agricultural production and thus to livelihood options and welfare.

Vapour Pressure Deficit

The expected influence of VPD was negative; the greater the deficit the lower expenditure would be. The results (Figures 17c and 17d) showed the relationship to be negative, except for in the extreme northeast and southwest regions of the country (beyond the green contours of Figure 17d).

VPD is a measure of the drying power of air and, where it has a significant influence, its coefficient against household expenditure is negative (i.e. the lower the VPD the higher the expenditure): in the north/central area of the country, above Lake Kyoga and the northwest shores of Lake Victoria. In the more arid northeast and the very humid southwest regions, where the sign of the coefficient is the reversed, VPD is not significant.

In the areas where the VPD coefficients are significant, the average VPD values are relatively low so, one might expect, not be limiting to welfare. It seems that in the drier areas of the northeast, where VPD is much higher and possibly more limiting to agricultural development and livelihood options, other variables are coming into play. NDVI in particular, is significant in these areas so presumably better accounts for the aridity. Also, the very different agricultural systems in central and northeast Uganda may be differentially affected by VPD (and NDVI)

Goat density

The expected influence of goat densities is ambiguous; in areas where goats are kept a positive effect might be expected (higher goat densities corresponding to higher expenditure), though in general goats are only kept in the more arid and isolated pastoral areas so are possibly indicative of lower average levels of welfare. The results (Figures 17e and 17f) showed a distinct northeast to southwest trend; with negative sign in the central region, and positive sign at either end of that trend (beyond the green contours of Figure 17f).

There are three distinct areas of significance in the goat coefficient: (i) in the arid, northeast pastoral areas, where goat densities are at their highest, and their influence on welfare is strongly positive; (ii) in a central band (northwest to southeast), again of relatively high goat density, where their influence is negative; and (iii) an area in the southern part of the country, where goat densities are lower, and their influence is again positive. The negative effects in the flatter more humid central regions may well reflect the variation in growing conditions within that zone; goats are likely to be raised in the drier areas less suited to cropping which would give rise to them being associated with lower welfare levels in these otherwise productive and relatively affluent regions. Goat density is positively related to expenditure in the more arid pastoral areas in the northeast: in these generally poorer areas goats are of great importance and are likely to reflect relative wealth. The positive influence in the temperate southwest is difficult to explain.

Cattle density

The expected influence of cattle density was positive, with higher densities giving rise to, or reflecting, greater wealth (expenditure). The results (Figures 17g and 17h) again showed a distinct northeast to southwest trend; with negative influence to the northeast of the green contour of Figure 17h), and positive influence southwest of it.

There are two areas in which the influence of cattle density was significant: (i) in the pastoral northeast, where the influence was negative, and (ii) in the southern part of the country, where the influence was positive. This latter result is easily explained: these are important cattle areas and dairy production, in particular, is prevalent. Why cattle densities should have a negative influence on expenditure in the northeast is difficult to fathom: one would expect a pattern similar to that of goats, whereby larger numbers reflect greater wealth among the pastoralists. It could be argued that too much should not be read into the coefficient map, however, since the parameter does not contribute strongly to the OLS model (c.f. Table 7 and Figure 14), and there are few areas where the GWR parameter is significant at the 5 percent level. Nonetheless, the pattern is intriguing and deserves further investigation.

Slope

The expected influence of slope was negative; steeper slopes corresponding to lower expenditure. The results (Figures 17i and 17j) indicated a strong east-west pattern; to the east of the zero contour (shown in green in Figure 17j there is a positive influence, whilst west of this contour the influence is negative.

The influence of slope is significant in three areas of Uganda: (i) the arid northeast, where its influence is positive; (ii) the northwest where its influence is negative, as expected; and (iii) the southwest of the country, where its influence again is negative, increasingly so from west to east and most strongly so close to the shores of Lake Victoria. Where it is significant, therefore, its influence is negative in the mixed farming areas, which is to be expected, since rough terrain hinders cultivation. Slope is less important in areas dominated by livestock, such as the northeast; the significant positive slope in this region therefore does not contradict the conclusions drawn from the other regions.

Population density

The expected influence of population density was positive; the greater the density of people the higher expenditure would be. The results (Figures 17k and 17l) revealed a strong north-south pattern; more positive in the north and less positive in the south of the country, but with no regions in which the influence of population density on expenditure was negative (there is no green, zero contour in Figure 17k).

The influence of population density is significant over most of the country, the exceptions being the southwest border, and a curiously-shaped wedge, fanning out to the east of Lake Kyoga. Both of these areas are where the population density coefficients are at their lowest values and, incidentally, they coincide with areas of steeply sloping land: Mount Elgon in the east and the Ruwenzori Mountains to the west. The population coefficient is more positive in the north than in the south, suggesting that the influence on expenditure of population density is less in areas of high rather than low population density. This may point to a saturation effect, in that there are diminishing returns to being near or in a high density area above a certain density threshold.

Travel time to markets

The expected influence of travel time to markets is negative; in general poor areas tend to be remote and higher expenditure would be expected in areas with good market access (with quick access to markets). The results (Figures 17m and 17n) showed bimodal, east-west trend; with negative sign in the more remote west and in the eastern parts of the country, beyond the green contours of Figure 20m, and positive sign within those contours, in the central and southwest parts of the country.

Travel time to markets was, perhaps surprisingly, one of the least influential parameters in the OLS regression model (c.f. Table 7 and Figure 11) and there are few areas where the GWR parameter is significant (Figure 17n), which makes interpretation difficult. There are three areas of significance: the eastern border of the country, where market access is poor and the influence of travel time is negative, as expected; (ii) a small area to the west, on the shores of Lake Albert, where market access is again generally poor and the influence of travel time is again negative, as expected; and (iii) a small area to the west of Lake Kyoga (only significant at the 10 percent level), where there is a surprising positive influence of travel time on expenditure. The patterns in the east and west suggest that increased access to markets for the more isolated regions would be beneficial. Interestingly though, this variable does not have a significant effect in all areas that are far from markets: the northwest, for example.