Validation of a Complex Spatial Model of the Food and Nutrition System in Zimbabwe

J A Wright, Institute of Ecology and Resource Management, University of Edinburgh, Agriculture Building, West Mains Road. Edinburgh EH9 3JG, Scotland, UK.
S W Gundry, Institute of Ecology and Resource Management, University of Edinburgh, Agriculture Building, West Mains Road, Edinburgh EH9 3JG, Scotland, UK.
A. Ferro-Luzzi, Istituto Nazionale della Nutrizione, via Ardeatina 546, 000178 Roma, Italia.
P.J. Hoyles, Istituto Nazionale della Nutrizione, via Ardeatina 546, 000178 Roma, Italia.

Abstract. This paper describes a method of validating a complex model of the food system in a district of Zimbabwe. The model will simulate patterns of malnutrition in all households within the district, using an artificial population created from census information and a sample survey. Model validation should require use of an independent data set not used in model construction. In many cases, such data are based on non-standard spatial units and may also be prone to bias. For the study district, the only nutritional data available are anthropometric measurements of children attending health centres, rather than district-wide household surveys. To use such data for validation, an attendance model has been developed that assesses the likelihood of children taking part in this health centre measurement programme. The probability of a given child being measured at a given facility is estimated based on the type of health facility, the characteristics of the child's household, the distance between the two, and the proximity of other equivalent types of facility. This relationship is estimated using logistic regression on field survey data and then applied to all of the households in the artificial population. The number of children attending each health facility can therefore be estimated, giving figures directly comparable to the aggregated health centre statistics. Incorporation of a module to simulate healthcare uptake means that aggregated model results can be validated for the youngest population cohort. Conditions in past years can be replicated within the model and output compared to historical records for different health centres across the district. This approach could also be applied to data collected at other types of fixed service point in rural areas, such as livestock diptanks, grain depots, or food aid distribution points. This work illustrates how secondary data, based on non-standard spatial units, can still be used for model validation.

1. INTRODUCTION

This paper addresses two problems commonly encountered in developing models of social behaviour. The first problem is the way in which the scale of observation influences the functional form of a given model. Perceived causes of phenomena at household level do not always coincide with those apparent in aggregate data for different social groups or regions. Several studies, for example, have illustrated how the apparent causes of disease clusters change with increasing levels of data aggregation [Waller and Turnbull, 1993; Schneider et al, 1993]. Such difficulties may also occur when analysing geographical patterns of under-nutrition.

The second problem concerns the issue of model validation that has been well documented in the past. Dent and Blackie [1979, pp. 100-102] have suggested four types of data that can be used for validating a model: historical data used in model construction; historical data not used in model construction; historical data collected since the model was constructed; and data explicitly collected for model validation. They suggest that the latter two types of data are more appropriate for validation, but admit that data collection solely for model validation is expensive. These two problems are addressed here by developing a model of health centre attendance based on detailed, household-level data. Aggregated data for each health centre – collected at a different scale - are then used to validate the model.

This approach to model validation is considered here in the light of an analysis of patterns of healthcare uptake in Zimbabwe. A simulation model is developed of participation by young children in a growth monitoring programme in the Buhera District in Manicaland Province. Factors affecting participation in the growth monitoring programme are identified on the basis of a household survey. This household-level model of attendance is then used to estimate the number of children weighed in the district's health centres. Simulation results are then compared to government statistics on the number of children attending each health centre, thereby validating the model using an 'unseen' data set.

1.1 The Growth Monitoring Programme in Zimbabwe

Since 1987, Zimbabwe has been operating a National Health Information System (NHIS) which collates information collected at health centres throughout the country [Tagwireyi and Greiner, 1994]. Part of the information collected through this system concerns the nutritional status of pre-school children, who have their age and weight recorded at health centres as part of a growth monitoring programme. The numbers of children who are weighed and the proportion who are underweight for their age are computerised under the NHIS, but details about the environment which the children come from are not recorded. This lack of information about the children's background makes it difficult to identify the causes of high levels of malnutrition at a particular clinic, without recourse

to community-based nutritional surveys of individuals. However, the NHIS growth monitoring data can still be used to validate a model derived from a community-based survey, *provided* that the NHIS data are adjusted to take account of different levels of participation in the weighing programme.

1.2 A Survey of Growth Monitoring Participation and Nutrition

A community-based field survey investigating patterns of healthcare use was undertaken in Buhera District in Manicaland province. Such a survey can be used both to identify the causes of under-nutrition and to assess the factors that influence participation in the growth monitoring programme described above. Buhera district consists of medium to low potential agricultural land, with annual rainfall varying from around 850 mm in the north to 550 mm in the south. Large rivers, the Save and the Nyazvidzi bound the district on two sides, and these barriers reduce the degree of interaction with neighbouring districts. Subsistence agricultural production is generally insufficient to meet requirements, so many adults work outside of the district in cities, mines or commercial farms, migrating on a seasonal basis and remitting funds back to the household. In terms of health facilities, the district is served by two fully equipped hospitals - Murambinda Mission hospital located in the north of the district, and Birchenough Bridge hospital in the south. Both hospitals lie close to tarmac roads, but there are only dirt roads in the remote, central part of the district. As well as the two main hospitals, the district contains nineteen smaller rural hospitals and clinics with more limited facilities. Schools and settlements in remoter parts of the district are also visited by mobile medical teams once a month, who perform immunisations, growth monitoring, and disseminate health information. Information about the socio-economic characteristics of households in the district is available from the 1992 Zimbabwean national census, broken down into 36 wards [Government of Zimbabwe, 1994].

354 households in 60 different villages within this district were selected to participate in the survey, based on a random sampling plan, stratified to capture variation in access to healthcare [Wright et al, 1996]. This represented a 1% sample of the total district population. The locations of participating villages and local health centres were recorded using a Global Positioning Systems (GPS) receiver and entered onto the GIS system.

2. A HOUSEHOLD MODEL OF GROWTH MONITORING UPTAKE

2.1 Methodology

In October 1995, adult carers of 284 children were asked to identify which clinics or hospitals (if any) their children had visited for growth monitoring. At the same time, the weights and heights of all household members participating in the survey were recorded. Initial investigation of the response to this question suggested that attendance at the two main hospitals followed a very different pattern from attendance at the other types of health centre. Two separate models of growth monitoring attendance were developed for these two types of facility, in which attendance at a particular health centre was expressed as a function of:

- the characteristics of the child and the child's family;
- the characteristics of the health centre;
- the distance from the child's home to the health centre
- and the extent to which other neighbouring facilities were able to offer similar facilities.

More formally, this relationship was expressed as a logistic regression equation, such that:

$$P = 1/(1 + e^{-Z}) \qquad (1)$$

(where P is the probability of a given child visiting a given health centre and e is the base of natural logarithms).

In this expression, Z is given by the equation:

$$Z = \beta_0 + \beta_1 X_1 ..... + \beta_n X_n + \varepsilon \qquad (2)$$

(where $X_1 ... X_n$ are terms representing the characteristics of the health centre, household and child, the distance from the child's home to the health centre, and competition from other health facilities; and ε is a residual error term).

This approach is derived from Rosenberg and Hanlon's [1996] study of healthcare uptake in Ontario, Canada. This study used logistic regression to model general uptake of health services based on demographic, income, and the health service environment, though attendance at specific facilities was not considered. The characteristics of the child and its household included in the analysis were restricted to those that could be obtained from the 1992 census for all the wards in Buhera District. Although this meant that some potentially useful predictors of formal healthcare participation were not considered, it meant that the results of the analysis could be applied to the whole district population by using the census information. Household characteristics considered in the analysis included the type of water source used, type of sanitation, and type of housing (a proxy measure for wealth).

Distances between surveyed villages and health centres were calculated in such a way that the effect of roads, rivers, and terrain on movement was explicitly incorporated. This was achieved through the use of a 'pushbroom' algorithm, which calculates distances from a 'difficulty of movement' map, rather than taking simple Euclidean distance [Eastman, 1989]. Slopes were calculated from a Digital Terrain Model of the district and converted to difficulty of movement based on human energy expenditure figures for different types of terrain [James and Schofield, 1990]. This was then combined with maps of difficulty of movement along roads and across rivers, derived from a series of interviews with local government staff from within the district (see Wright et al [1996] for a more comprehensive discussion of this procedure).

The extent to which other neighbouring health facilities influenced attendance at a given health centre was represented through a 'competition' variable. Competition was represented by subtracting the distance to the nearest

health facility capable of providing the same level of healthcare from the distance to the particular health facility being considered. The value of this competition variable was therefore zero for the nearest facility to a given village, and increased for facilities that were further away. Since the number of possible combinations of independent variables was large, a stepwise regression technique was used to identify the combination of variables that best explained the observed pattern of clinic attendance.

2.2 Household-Level Model Results

Table 1 shows the results of the stepwise logistic regression analysis for hospitals and rural hospitals/clinics. In the case of hospitals, growth monitoring participation was related to the type of water source used by households and to the distance between the survey village and the hospital. The probability of visiting a hospital declined with distance, whilst households using unsafe water sources (such as rivers or dams) were less likely to take their children for growth monitoring. For the smaller rural hospitals and clinics, the probability of growth monitoring attendance was related solely to the 'competition' variable discussed earlier. The fact that the 'competition' variable proved important for clinics and rural hospitals but not for the main hospitals may be related to the much lower density of main hospitals. Because the main hospitals are so far apart, it may be that they never effectively compete for patients - unlike the clinics and rural hospitals.

Variable	Type of facility	Sign:	Significance
Distance	Hospital	Negative	**
Water Access	Hospital		*
Constant	Hospital	Positive	*
'Competition'	Clinic	Negative	**
Constant	Clinic	Positive	**

Table 1: Results of Logistic Regression Analysis for Hospitals and Clinics (** = significant at the 99% level; * = significant at the 95% level. The hospital model Chi-square was 128.0 which was significant at the 99.9% level, and the clinic model Chi-square was 163.0, also significant at the 99.9% level[1]).

3. SIMULATION OF HEALTH CENTRE ATTENDANCE

In order to validate this relationship, these household-level findings were used as the basis for a simulation that estimated attendance at health centres throughout Buhera. Spatial relationships were important in the model, being represented by the distance and competition terms discussed in Section 2. However, information about the distribution of population was limited. The total number of children under 5 years living in each of the district's 36 wards was available from the August 1992 census [Government of Zimbabwe, 1994]. In addition, the locations of three types of major settlements were available from the local government authorities within the district: growth points (towns with electricity and telephone services), business centres (large villages), and rural service centres (smaller villages). Older, more detailed settlement maps dating from the period of the Rhodesian Unilateral Declaration of Independence were found to be out of date, confirming the dynamic nature of human settlements in the district noted elsewhere [Campbell et al, 1989]. The under-5 population of a number of rural service centres and business centres was estimated based on a rapid rural appraisal during 1995. In order to create a population density map, the mean number of children estimated as resident was assigned to each rural service centre and business centre in the district. The residual under-5 ward population not accounted for by these major settlements was distributed amongst villages randomly located within the ward.

The model described in Section 2 was then applied to every child in the district. The probability of attendance for each individual was calculated based on type of water source and distance for hospitals, and on the competition variable for clinics. The mean number of children attending each health facility was then calculated by multiplying these probabilities by the number of children at each point in the district.

The questionnaire survey had asked whether or not children under 5 years old had attended health centres, without considering the frequency of visits or the period of time during which the visits took place. Consequently, two adjustments were made so that monthly attendance at health centres could be estimated using the method described above. The questionnaire asked about attendance during a child's lifetime, rather than over a specified period. To compensate for this, the estimated number of visits from the simulation was divided by the average age in months of children in the under-5 age cohort (29 months) to convert to a monthly figure. In addition, the questionnaire considered which health centre a given child had attended, but not the frequency of visits to that health centre. Government of Zimbabwe [1988] note from an extensive national survey of growth monitoring attendance that children on average make four visits in their first year of life, two in their second, and one in their third. The mean number of visits declined further in the fourth and fifth years, as fewer visits to health centres were needed for immunisation. To incorporate the frequency of visits, the number of simulated visits was then multiplied by the mean frequency of visits by children under 5 (6.4 visits). This gave an estimated mean monthly attendance figure that could be directly compared to the data from the NHIS.

To validate these results, the number of children attending growth monitoring at each health centre per month, averaged over the period January 1990 to September 1995 was then calculated from the NHIS data. Simulation output was then regressed on these actual attendance figures. Following the approach adopted by Kleijnen and Van Groenendaal [1992], an intercept term significantly different from zero was taken as evidence for rejecting the model. Similarly, a slope coefficient significantly different from one would also be taken as evidence for rejecting the model. These two assumptions can be tested simultaneously through the use of an F-statistic [Harrison, 1996], as given by the formula:

[1] The model Chi-square is equivalent to the F test for ordinary linear regression.

$$\frac{(n-2)\{nb_1^2 + 2nb_1(b_2 - 1) + \Sigma x_t^2(b_2 - 1)^2\}}{2ns^2}$$

where n is the sample size, s^2 is the residual variance, b_1 is the regression intercept coefficient, b_2 is the regression slope coefficient, and x_t is the tth observation of the real system output.

4. RESULTS AND DISCUSSION

Table 2 shows the results of regressing the real-world data on the simulated numbers of visits. The household-level model passed the validation test, both when all health centres were included and when health centres likely to be visited by children from outside the district were excluded. The regression, which was based on 21 health centres, had an adjusted R^2 of 0.528. When T-tests were applied to the results in Table 2, the constant term was found to be significantly different at the 99% level from zero and the slope co-efficient was found to be significantly different from one at the 99% level (although it was significantly different from zero). The model also failed a simultaneous F-test of goodness of fit when this was applied to the regression statistics.

Term	*Coefficients*	*Standard Error*	*Significance*
Intercept	112.7056	22.87887	99.5%
Slope	0.332552	0.068831	99.5%

Table 2: Results of regressing real world data on model output.

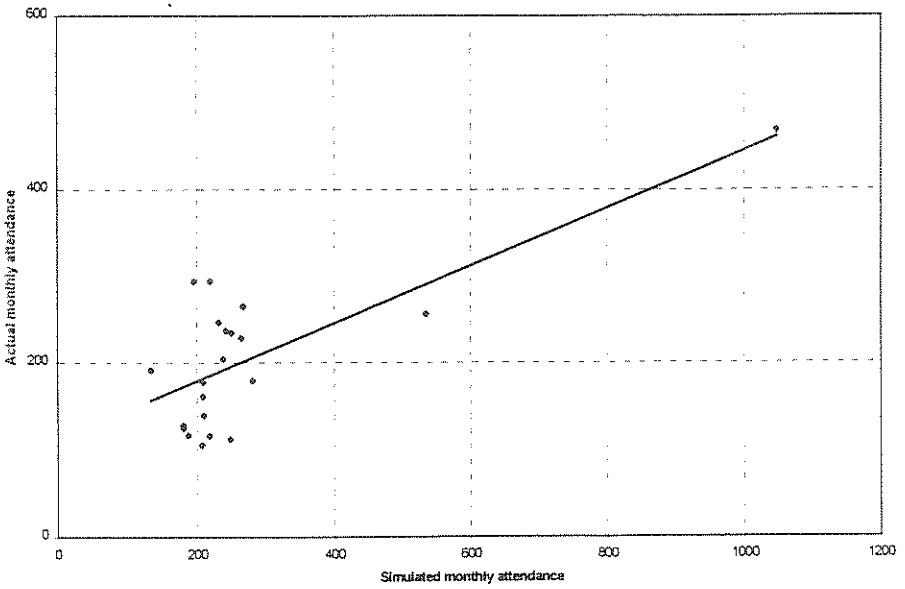

Figure 1: Scatterplot of actual mean number of children under 5 attending growth monitoring per month for 21 health centres and the simulated number attending (the fitted regression line is also shown).

Figure 1 shows the relationship between actual monthly clinic visits and the clinic visits simulated by the model. Although the simulation model successfully predicts a major difference between the number of attendees at the principal district hospital, Murambinda, and the other health centres, it is less successful at discriminating differences between attendance at other clinics.

Several sources of variation that reduce the goodness of fit between the model output and actual attendance figures can be identified:

- Variation associated with the estimation of attendance behaviour from the household survey data;
- Variation associated with the creation of the population density map for the district;
- Variation associated with the rescaling of the simulation results to take account of the frequency of visits by children;
- The confounding influence of visits by children from outside the district, although this is reduced for much of the area by the natural barriers of the Save and Nyazvidzi rivers;
- Transcription, data entry, clerical, and other errors associated with the collation of the NHIS actual attendance figures.

Given these five different sources of variation, the fact that the simulation exercise failed the simultaneous F-test is unsurprising. However, the fact that the simulation model tended to over-estimate attendance suggests that the rescaling adjustment to account for the frequency of health centre use is one of the major sources of error in the model. A revised questionnaire design, asking about frequency of

attendance over a specified period, would resolve this difficulty.

In addition, several authors have argued that the simultaneous F-test is too rigorous for agricultural and socio-economic model development. Some have questioned its use with large validation data sets [Thornton and Hansen, 1996], whilst Harrison [1990: p. 183] has suggested that instead: 'descriptive statistics and subjective tests be used to build up confidence in a model as it proceeds through a number of prototypes'. The scatterplot presented in Figure 1, and the significant, positive slope coefficient identified in the regression exercise both suggest that output from this early prototype of a clinic attendance type is well correlated with actual attendance figures.

5. CONCLUSION

Health centre-level data were used to validate a household-level model. The household-level model failed a regression-based validation test, but was able to distinguish between poorly attended clinics and well-attended hospitals. The attractiveness of hospitals to patients is thus borne out both by the household survey and by health statistics collected by government. Although the model failed a simultaneous F-test, more descriptive investigation of its output suggested that this was correlated with actual attendance data.

Further refinement of this attendance model will now be made, so that it can eventually be combined with a model of the causes of poor nutrition. The household survey described here will be used to develop a model of under-nutrition prevalence in children under 5 years, which uses a rule-base to determine the nutritional status of all members of the household [Gundry et al, 1997]. Using the attendance simulation described here, such a model can then be tested against data for children measured at health centres. In addition, the same technique applied here can be used to identify areas where the probability of healthcare uptake is low so that suitable locations can be identified for potential new facilities.

Acknowledgements

This research was funded by the Commission of the European Communities DGXII, Science and Technology for Developing Countries Programme, contract reference TS3*-CT92-0048. The authors wish to thank Mrs. J. Nyatsanza of the University of Zimbabwe, staff at Buhera District Development Fund, Dr. Glenshaw and Mr. Jephres Masocha of Murambinda Hospital for their help in data preparation and Prabhat Vazé of the Department of Economics, University of Edinburgh for comments on this work.

References

Campbell, B.M., du Toit, R.F., and Attwell, C.A., *The Save study: relationships between the environment and basic needs satisfaction in the Save catchment, Zimbabwe.* University of Zimbabwe, 119 pp., Harare, 1989.

Dent, J.B., and Blackie, M.J., *Systems Simulation in Agriculture.* Applied Science, 180 pp., London, 1979.

Eastman, R., Pushbroom algorithms for calculating distances in raster grids, *Proceedings, AUTOCARTO* 9: 288-297, 1989

Government of Zimbabwe, *Primary health care / maternal and child health / expanded programme on immunisation surveys.* Ministry of Health, Harare, 1988.

Government of Zimbabwe, *Census Provincial Profile: Manicaland.* Central Statistical Office, 147 pp, Harare, 1994.

Gundry, S.W., Wright, J.A., and Ferro-Luzzi, A., Simulating the food and nutrition system in rural Zimbabwe to support targeting of emergency food aid. MODSIM 97 International Congress on Modelling and Simulation, Hobart, Australia, 8-11th December, 1997.

Harrison, S.R., Regression of a model on real-system output: an invalid test of model validity, *Agricultural Systems* 34: 183-190, 1990.

James, W.P.T., and Schofield, E.C., *Human Energy Requirements: a Manual for Planners and Nutritionists.* Oxford University Press, 172 pp., Oxford, 1990.

Kleijnen, J.P.C., and Van Groenendaal, W., *Simulation: a Statistical Perspective.* Wiley, 187 pp, Chichester.

Rosenberg, M.W., and Hanlon, N.T., Access and utilization: a continuum of health service environments, *Social Science and Medicine* 43 (6): 975-983, 1996.

Schneider, D., Greenberg, M.R., Donaldson, M.H., and Choi, D., Cancer clusters: the importance of monitoring multiple geographic scales. *Social Science and Medicine*, 37: 753-759, 1993.

Tagwireyi, J. and Greiner, T., *Nutrition in Zimbabwe: an Update.* The World Bank, 127 pp, Washington D.C., 1994.

Thornton, M., and Hansen, J.W., A note on regressing real-world data on model output. *Agricultural Systems* 50: 411-414, 1996.

Waller, L.A., and Turnbull, B.W., The effects of scale on tests for disease clustering. *Statistics in Medicine*, 12: 1869-1884.

Wright, J., Hoyles, P.J., Makombe, G., Gundry, S.W., Mudimu, G. and Nyatsanza, J., The use of Geographical Information Systems for household survey sample stratification: an application from Zimbabwe in nutrition and food systems research. Development Studies Association Conference, University of Reading, 18-20th September, 1996.