Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 6557-6564.

Modeling forest conversion to pasture at the Arch of deforestation in the Brazilian Amazon through linear regression technique

Fabiano Costa de Almeida [1,2]
John Mauricio Arenas [2]
Leonardo Marini Pereira [2]
Ana Paula Dutra de Aguiar [2]
Corina da Costa Freitas [2]

[1] Diretoria de Serviço Geográfico - DSG
QGEx - Bloco "F" - SMU - 70630-901 - Brasília - DF, Brazil
fcdealmeida@dpi.inpe.br

[2] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brazil
{john, leonardo}@dsr.inpe.br, {anapaula, corina}@dpi.inpe.br

Abstract. The aim of this paper is to determine a linear regression model that explains the forest conversion to pasture at the Arch of deforestation in the Brazilian Amazon. Data from regular cells (25 x 25 Km²) of an initial set of nineteen explanatory variables were processed and, after mathematical transformations and statistical analysis, a linear regression model with eleven variables was obtained. Results show that connection to São Paulo (national market), distance to railroads, percentage of protected area and population density are the most important factors for the model – greatest values of beta coefficients. Finally, it is noted that, when euclidean distance to nearest road (a twentieth explanatory variable that had not been considered previously) is included, population density variable is replaced with the former in the subset of the highlighted beta values without even remaining in the model.

Keywords: LUCC, deforestation, Brazilian Amazon, linear regression model.

1. Introduction

The Amazon Forest remained "unharmed" until the commonly called "era of deforestation" initiates with the instatement of Trans-Amazon Road in 1970 (Fearnside, 2005). Migration on Legal Amazon began in context of national integration policy project which included poles of development, land appropriation for agriculture projects and agrarian reform, mining and more recently, grain exportation. Since 1970, migration process expanded and millions of hectares of forestall lands were occupied for range landing implantation, colonization and agrarian reform policy application (Alves, 2001).

At this intricate context, decision makers need to understand the determinant factors and the anticipation of possible situations and scenarios. The task is the identification between proximate causes and underlying driven forces for not giving superlative importance of a single factor (Aguiar, 2006; Veldkamp and Lambin, 2001).

Several approaches on Land Use and Land Cover Change (LUCC) Modeling have been done with different emphasis, goals, perspectives and data (Aguiar, 2006; Hietel et al., 2006; Irwin and Geogegan, 2001; Lambin et al., 2001; Veldkamp and Fresco, 1996). Models have been developed based just on social-economic theory, or on spatial context; involving both approaches and just few of them exploring intra-regional differences (Aguiar, 2006).

This work is aimed at presenting a model that explains the forest conversion to pasture in the Brazilian Amazon by exploring regional scale data from regular cells (25 x 25 Km²) at the Arch of deforestation through linear regression technique.

2. Study Area and Data

The study area for application of linear regression model is the Arch of deforestation or the densely populated Arch called by Becker (2005) of one of the three macro-regions in Amazon (**Figure 1**).

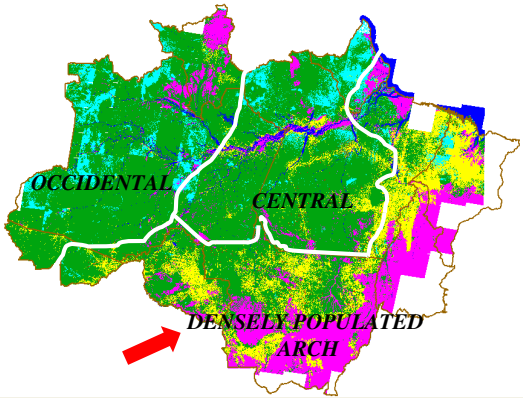

Figure 1 – Study area: densely populated Arch [source: Aguiar (2006)].

Aguiar (2006) collected and processed data into 2026 regular cells in fine scale resolution (25 x 25 km²) using CLUE framework modeling. Only a subset of variables provided for that author was used in this work and the explanatory ones are divided into categories as shown on **Table 1**.

Table 1 – Categories of explanatory variables [source: Aguiar (2006)].

Category	Cellular Database Variable	Description	Source
Accessibility to Markets	Dist_Rivers	Euclidean distance to nearest large river [Km]	IBGE
	Dist_Railroads	Euclidean distance to nearest railroad [Km]	IBGE
	Conn_sp_inv_p	Connection to SP (national market) through the road network considering the type of road	IBGE
	Conn_ne_inv_p	Connection to NE (national market) through the road network considering the type of road	IBGE
Economical Attractiveness	Dist_Wood	Euclidean distance to poles of timber production [Km]	IBAMA
	Dist_Mineral	Euclidean distance to all types of mineral deposits [Km]	CPRM
Demographical	Pop_Dens_96	Population density in 1996	IBGE
	Pop_Tot_Var_81_91	Total population variation between 1981-1991	IBGE
Technological	Tech_Tractor	Number of tractor per number of property owners	IBGE
	Tech_Fertilizer	Number of fertilized properties per number of property owners	IBGE
Agrarian Structure	Agr_Area_Small	Percentage of small and large properties in terms of municipalities area [% of cell area]	IBGE
	Agr_Area_Large		
	Agr_nr_Small	Percentage of small and large properties in terms of number of properties in the municipalities [% of cell area]	IBGE
	Agr_nr_Large		
Political	Settl_nfamilies_70_99	Number of settled families until 1999	INCRA
	Prot_all	Percentage of protected area (any type) [% of cell area]	IBAMA/FUNAI
Environmental	Soil_fert_B1	Percentage of soils of high and medium fertility [% of cell area]	IBGE
	Clima_humi_min_3_ave	Humidity mean (May, Jun, Jul, Aug) [%]	INMET
	Clima_humi_min_3_ave	Total precipitation (May, Jun, Jul, Aug)	INMET

The pattern of distribution of dependent variable Luc_Past (percentage of pasture area in 1996/1997) is shown on **Figure 2**.

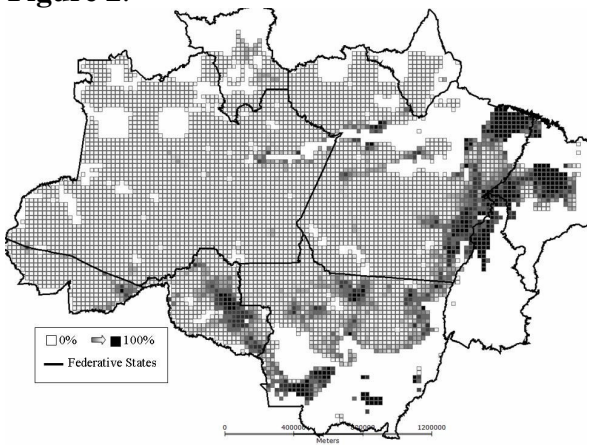

Figure 2 – Pattern of distribution of dependent variable [source: Aguiar (2006)].

3. Building the linear regression model (LRM)

Neter et al. (1996) present the fluxogram (**Figure 3**) that was used to build the linear regression model in this study.

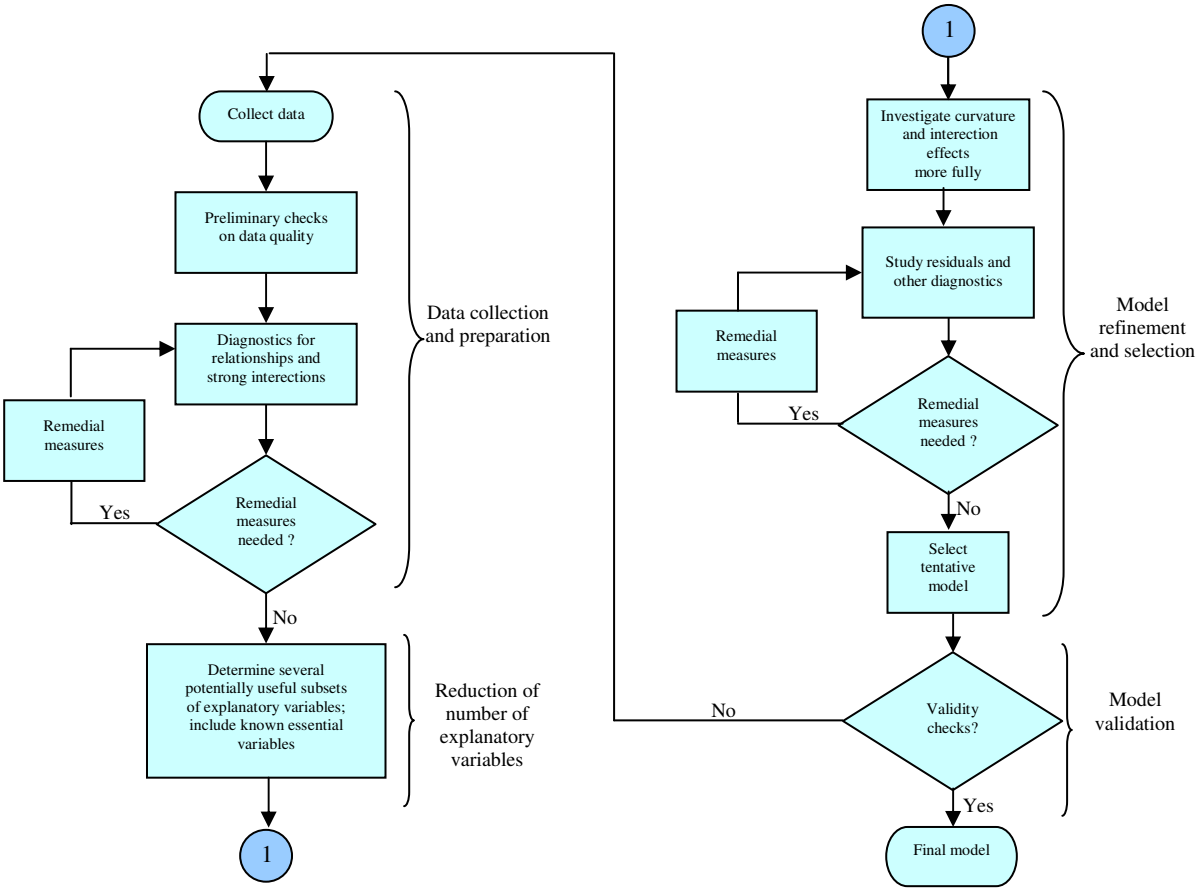

Figure 3 – Fluxogram for building regression model [source: Neter et al. (1996)].

The dependent variable and the nineteen predictor ones were compared into a matrix correlation table and a scatterplot matrix graph for finding out correlations and relationships among variables. They indicated the necessity of applying transformations over the variables.

The Shapiro-Wilk test indicated non-normality of dependent variable. It was expected due to the large number of null value cases (**Figure 4**).

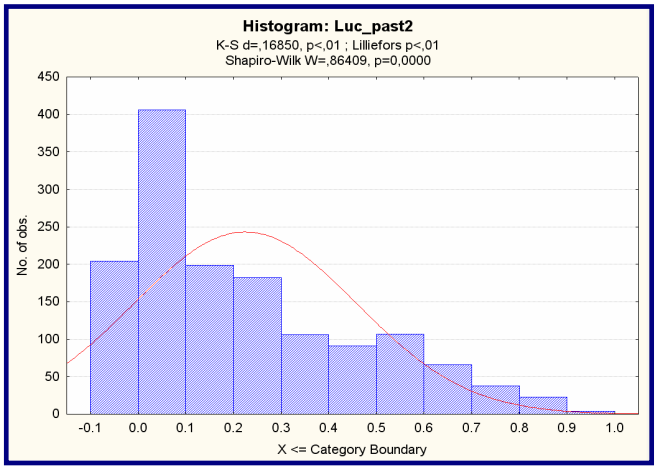

Figure 4 – Exploratory plot of dependent variable.

The analysis of residuals suggested the linearity of regression function, nonconstancy of error variance, presence of outliers, independence and normality of error terms.

Logarithmic and square root functions were applied on dependent variable in order to find out better correlation and minimize the non-constant tendency of its variance. The same transformations were applied on explanatory variables to improve the analysis and relational comprehension based on Pearson coefficient.

The square root transformation of the dependent variable ($Y^{0.5}$) was the best one and the functions for the explanatory ones are shown on **Table 2**.

Table 2 – Transformations applied on explanatory variables.

Variable	Function	Variable	Function
Dist_Rivers	None	**Agr_Area_Small**	Logarithmic
Dist_Railroads	None	**Agr_Area_Large**	None
Conn_sp_inv_p	None	**Agr_nr_Small**	Logarithmic
Conn_ne_inv_p	Square Root	**Agr_nr_Large**	None
Dist_Wood	Logarithmic	**Settl_nfamilies_70_99**	Square Root
Dist_Mineral	Logarithmic	**Prot_all**	None
Pop_Dens_96	Logarithmic	**Soil_fert_B1**	None
Pop_Tot_Var_81_91	Logarithmic	**Clima_humi_min_3_ave**	None
Tech_Tractor	Logarithmic	**Clima_precip_min_3_ave**	Logarithmic
Tech_Fertilizer	Logarithmic		

The log(Tech_Tractor) variable is highly correlated with several explanatory variables. It was decided to exclude this one in order to avoid ill-conditioning during the mathematical regression procedure and also to reduce information ambiguity.

Three variables (log(Pop_Tot_Var_81_91), log(Clima_precip_min_3_ave) and Agr_Area_Large) were also excluded because of the high correlation with other one in their categories (0.79 with log (Pop_Dens_96), 0.76 with Clima_humi_min_3_ave and -0.89 with log(Agr_Area_Small), respectively).

The variables log(Agr_nr_Small) and Agr_nr_Large were eliminated due the negligible correlation value with the dependent one (0.03 and 0.06, respectively). These variables become useless for explanatory variables in the LUCC model.

Analysis of p-value by a t-test suggested exclusion of log(Agr_Area_Small) for α =5% and Dist_Rivers for α=1%. A forward stepwise procedure suggested the same variables to purge by the last steps applied over the correlation linear model.

An F-test was conduced to conclude if both variables are non-significant at the same time. The test was conduced considering a full model with 13 variables and a reduced model with 11 variables. The test concluded that both are non-significant for α=1%, as suggested the t-test and the forward stepwise procedure.

As a complementary and decisive method, the all-possible-regressions procedure was implemented, which considered all possible subsets in the set of potential X variables letting identify and inspect just the "best" subsets according the individual criterion for each approach.

The R^2 (**Figure 5(a)**) and adjusted R^2 (**Figure 5(b)**) criteria reached the 0.69 as its top determination coefficient value with a medium set of possible variables involved (7 to 13) for being implemented as the chosen model between the ranges of 0.66 to 0.69.

However, the Cp value reduced significantly the possible set of variables combinations that could be "suitable" for the model. The chosen set was for 11 variables (n + 1 is equal of number of parameters, using as criterion of choosing), the Cp value was 11.5, which was the lowest one for all subsets combinations (**Figure 5(c)**).

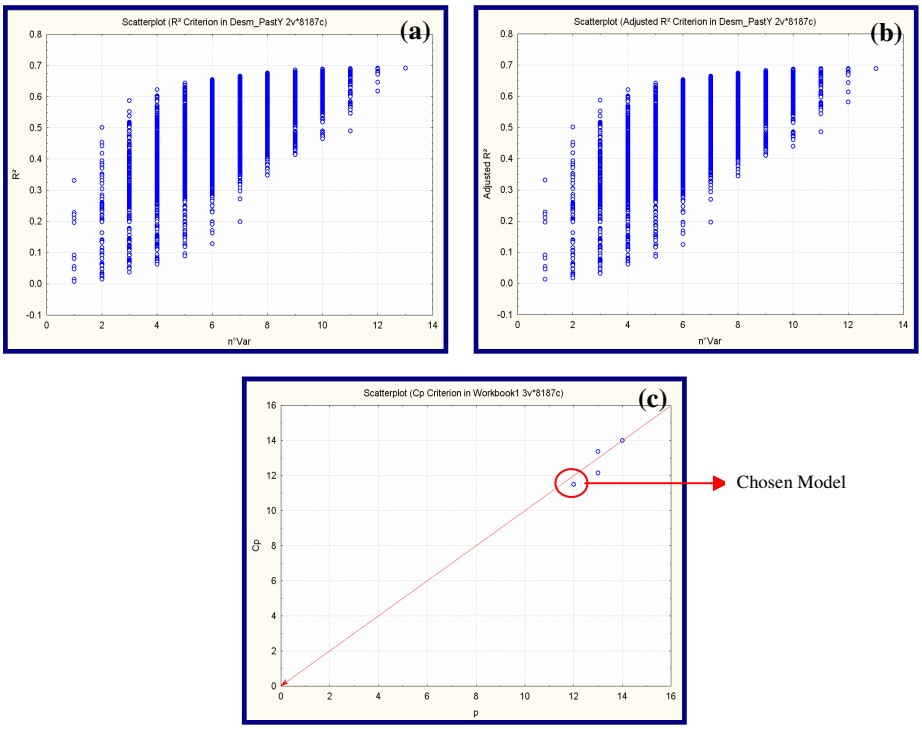

Figure 5 – R², adjusted R² and Cp criteria.

The correlations among the variables of the chosen model are presented on **Table 3** and its regression summary, on **Table 4**.

Table 3 – Correlation matrix for the chosen model.

	LOG (Pop_dens_96)	LOG (Tech_Fertilizer+1)	Sqrt(Setl_nF)	Prot_all1	LOG (Dist_wood)	LOG (Dist_Min)	Dist_Railroads	Conn_SP_Inv_P	Sqrt(Conn_ne)	Clima_Humi_Min_3_Avi	Soils_Fert_B1	Y^0,5
LOG(Pop_dens_96)	**1.000**	0.022	0.353	-0.122	-0.142	-0.264	-0.443	0.120	0.480	-0.001	0.194	0.479
LOG(Tech_Fertilizer+1)	0.022	**1.000**	-0.120	-0.085	-0.105	0.233	0.009	0.320	0.025	-0.267	0.173	0.038
Sqrt(Setl_nF)	0.353	-0.120	**1.000**	-0.104	0.030	-0.107	-0.266	0.063	0.243	-0.053	0.031	0.286
Prot_all1	-0.122	-0.085	-0.104	**1.000**	0.100	0.052	0.093	-0.209	-0.213	0.076	-0.098	-0.445
LOG(Dist_wood)	-0.142	-0.105	0.030	0.100	**1.000**	0.179	-0.212	0.103	0.136	-0.386	-0.078	-0.088
LOG(Dist_Min)	-0.264	0.233	-0.107	0.052	0.179	**1.000**	-0.013	0.025	0.004	-0.338	0.052	-0.218
Dist_Railroads	-0.443	0.009	-0.266	0.093	-0.212	-0.013	**1.000**	-0.199	-0.548	0.348	-0.026	-0.458
Conn_SP_Inv_P	0.120	0.320	0.063	-0.209	0.103	0.025	-0.199	**1.000**	0.528	-0.556	0.167	0.576
Sqrt(Conn_ne)	0.480	0.025	0.243	-0.213	0.136	0.004	-0.548	0.528	**1.000**	-0.343	0.110	0.472
Clima_Humi_Min_3_Avi	-0.001	-0.267	-0.053	0.076	-0.386	-0.338	0.348	-0.556	-0.343	**1.000**	-0.069	-0.304
Soils_Fert_B1	0.194	0.173	0.031	-0.098	-0.078	0.052	-0.026	0.167	0.110	-0.069	**1.000**	0.237
Y^0,5	0.479	0.038	0.286	-0.445	-0.088	-0.218	-0.458	0.576	0.472	-0.304	0.237	**1.000**

Table 4 – Regression Summary

R= ,83597043 R²= ,69884656 Adjusted R²= ,69650378

F(11,1414)=298,30 p<0,0000 Std.Error of estimate: ,15301

	Beta	Std.Err.	B	Std.Err.	t(1411)	p-level
Intercept			1.2604	0.149243	8.4452	0.000000
LOG(Pop_dens_96)	0.245854	0.020014	0.1133	0.009224	12.2840	0.000000
LOG(Tech_Fertilizer+1)	-0.171948	0.016864	-0.4597	0.045087	-10.1963	0.000000
Sqrt(Setl_nF)	0.064787	0.015946	0.0032	0.000777	4.0629	0.000051
Prot_all1	-0.287469	0.015261	-0.2540	0.013484	-18.8373	0.000000
LOG(Dist_wood)	-0.136202	0.016898	-0.1080	0.013394	-8.0604	0.000000
LOG(Dist_Min)	-0.110961	0.017136	-0.0683	0.010544	-6.4754	0.000000
Dist_Railroads	-0.300484	0.019504	0.0000	0.000000	-15.4066	0.000000
Conn_SP_Inv_P	0.535703	0.021975	153.9667	6.315700	24.3784	0.000000
Sqrt(Conn_ne)	-0.181267	0.022468	-3.5292	0.437442	-8.0679	0.000000
Clima_Humi_Min_3_Avi	-0.066901	0.021467	-0.0034	0.001106	-3.1165	0.001867
Soils_Fert_B1	0.102160	0.015334	0.0884	0.013268	6.6623	0.000000

The Shapiro-Wilk test was applied for normal analysis of residuals accepting the null hypothesis for α = 1% (**Figure 6(a)**). The normal probability plot (**Figure 6(b)**) shows high normality based on coefficient of correlation.

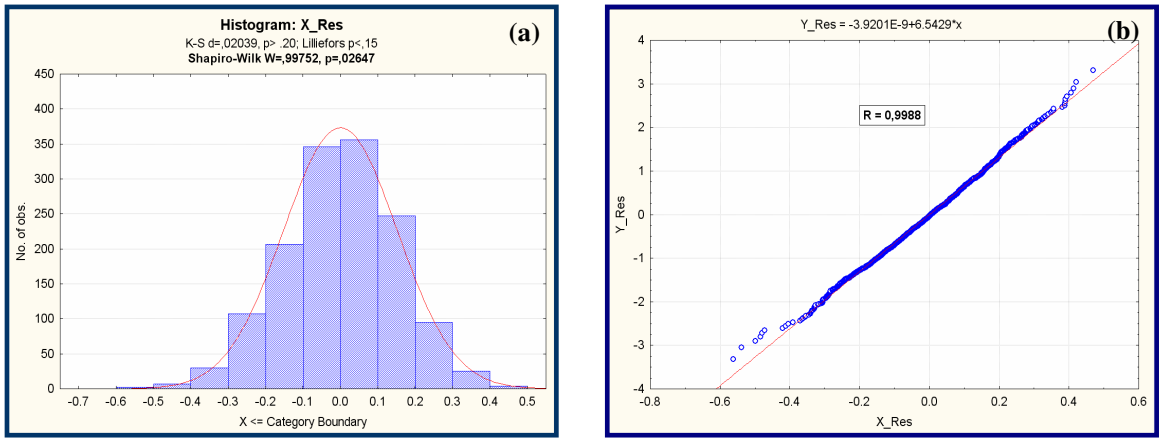

Figure 6 – Normal analysis of residuals.

Modified Levene test indicated homocedasticity for α = 1%.

Outlying cases did not affect the whole model according to Cook's distance measure and only 4.9% of cases were identified as influent values on themselves by DFFITS values.

Since the predictor variables are correlated among themselves, multicollinearity takes place in the model. The effects of this situation are various: inhibit the ability of obtaining good fit, estimated regression coefficients individually may not be statistically significant and simple inference of a specific variable while the other ones are held constant it is no longer applicable.

Then, a metric measure called variance inflation factor (VIF) was applied. Its mean value is 1.628546, concluding that no multicollinearity affects the model (mean VIF is not considerably larger than 1).

Among the several methods used for validation a model, mean squared prediction error (MSPR) was the selected one. Thirty per cent of original data (600 of 2026 cases) were previously separated for this procedure. Since MSPR value (0.023568) is fairly close to mean square error (0.023411), the selected model it is not seriously biased and gives an appropriate indication of the predictive ability of the model.

The final LRM obtained is shown on **Equation 1**.

$$Y^{0.5} = 1.2604 + 0.1133*\log X_1 - 0.4597*\log(X_2+1) + 0.0032*X_3^{0.5} - 0.2540*X_4 - 0.1080*\log X_5 - 0.0683*\log X_6 - 2.7*10^{-7}*X_7 + 153.9667*X_8 - 3.5292*X_9^{0.5} - 0.0034*X_{10} + 0.0884*X_{11} + \xi. \quad \textbf{(1)}$$

Where:

Y = Luc_Past;	X_5 = Dist_Wood;	X_{10} = Clima_humi_min_3_ave;
X_1 = Pop_Dens_96;	X_6 = Dist_Mineral;	X_{11} = Soil_fert_B1;
X_2 = Tech_Fertilizer;	X_7 = Dist_Railroads;	ξ = Random error term.
X_3 = Settl_nfamilies_70_99;	X_8 = Conn_sp_inv_p;	
X_4 = Prot_all;	X_9 = Conn_ne_inv_p;	

The most important explanatory variables for the model are highlighted by the beta values. They are: Conn_sp_inv_p, Dist_Railroads, Prot_all and Pop_Dens_96.

In order to verify if the euclidean distance to nearest road influences the model, the variable Dist_Roads – a twentieth explanatory variable that had not been considered previously – was incorporated to the data set.

All the steps were replicated and the final model has the same explanatory variables, excepted for the replacing log(Pop_Dens_96) with the square root of Dist_Road – including its place in the highlighted beta values set (**Table 5**).

Table 5 – Regression Summary (considering the distance to nearest road variable)

R= ,83551737 R²= ,69808927 Adjusted R²= ,69574060
F(11,1414)=297,23 p<0,0000 Std.Error of estimate: ,15320

	Beta	Std.Err.	B	Std.Err.	t(1411)	p-level
Intercept			1.4962	0.147925	10.1145	0.000000
LOG(Tech_Fertilizer+1)	-0.132716	0.016746	-0.3548	0.044772	-7.9252	0.000000
Sqrt(Setl_nF)	0.071190	0.015858	0.0035	0.000773	4.4893	0.000008
Prot_all1	-0.274234	0.015286	-0.2423	0.013507	-17.9397	0.000000
LOG(Dist_wood)	-0.111728	0.017424	-0.0886	0.013811	-6.4123	0.000000
LOG(Dist_Min)	-0.145645	0.016599	-0.0896	0.010213	-8.7743	0.000000
Dist_Railroads	-0.317844	0.019216	0.0000	0.000000	-16.5408	0.000000
Sqrt(Dist_Roads)	-0.228890	0.018880	-0.0007	0.000056	-12.1231	0.000000
Conn_SP_Inv_P	0.428526	0.022398	123.1628	6.437290	19.1327	0.000000
Sqrt(Conn_ne)	-0.120569	0.021221	-2.3475	0.413177	-5.6815	0.000000
Clima_Humi_Min_3_Avi	-0.072079	0.021515	-0.0037	0.001108	-3.3502	0.000829
Soils_Fert_B1	0.112589	0.015214	0.0974	0.013163	7.4005	0.000000

This is an important statement because distance from roads could be measured easier by remote sensing users/researchers than demographical data collection.

In spite of data set are spatially distributed, the approach used in this work did not consider the spatial effects. In order to provide a visual analysis, a map is presented containing residuals cell frame (**Figure 7**).

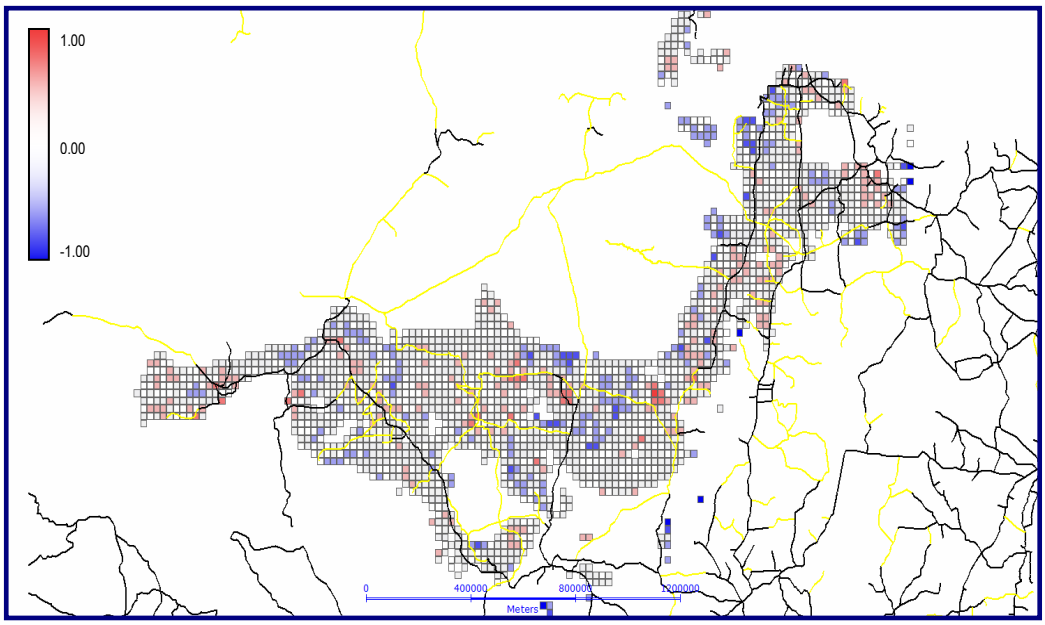

Figure 7 – Spatial distribution for standardized residuals.

The global Moran index of 0.38 was calculated, indicating medium spatial dependence.

4. Conclusions

Data from 2026 regular cells (25 x 25 Km²) of an initial set of nineteen explanatory variables were processed and refinements procedures produced a linear regression model with eleven explanatory variables. The most important ones were highlighted by the beta values: connection to São Paulo, distance to railroads, percentage of protected area and population density.

This result ratifies the heterogeneity of factors that should explain the forest conversion to pasture at the Arch in the Brazilian Amazon in the last decades. While accessibility to markets motivates the deforestation, the political activity of creating protected areas appears as an

effective method to avoid this situation. These factors are followed by a demographic one (population density) which is able to make deforestation worse.

Incorporating the euclidean distance to nearest road variable to the data set, the final model has the same explanatory variables. Except for the replacing population density variable with the former in the final model – including its place in the highlighted beta values.

The conclusions at this point should be the same, but there is an extra advantage: distance from roads could be measured easier by remote sensing users/researchers than demographical data collection.

Although the approach used in this work did not consider the spatial effects, the distribution for standardized residuals in the map suggests spatial dependence which is confirmed by the global Moran index as a medium one.

References

Aguiar, A. P. D. **Modeling land use change in the Brazilian Amazon: exploring Intra-regional heterogeneity**. 2006. Remote Sensing Doctoral Thesis. Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2006. In Press.

Alves, D.S. O processo de desmatamento na Amazônia. **Parcerias Estratégicas**, n.12, p.259-275. 2001.

Becker, B. K. Geopolítica da Amazônia. **Estudos Avançados**, v. 19, n. 53, 2005. Available from: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142005000100005&lng=en&nrm=iso>. Access on: 13 Nov 2006.

Fearnside, P. M. Deforestation in Brazilian Amazonia: history, rates and consequences. **Conservation Biology**, v. 19, n. 3, p. 680-688. 2005.

Hietel E.; Waldhardt R.; Otte A. Statistical modeling of land-cover changes based on key socio-economic indicators. **Ecological Economics**. 2006. In Press. Avaiable from: <http://www.sciencedirect.com/science/journal/09218009>. Access on: 13 Nov 2006.

Irwin, E.; Geoghegan, J. Theory, data, methods: developing spatially-explicit economic models of land use change. **Agriculture, Ecosystems and Environment**, v. 85, p. 7-24, 2001.

Lambin, E. F.; Turner, B. L. I.; Geist, H. J.; Agbola, S. B.; Angelsen, A.; Bruce; J.W.; Coomes, O.; Dirzo, R.; Fischer, G.; Folke, C.; George, P. S.; Homewodd, K.; Imbernon, J.; Leemans, R.; Li, X.; Moran, E. F.; Mortimore, M.; Ramakrishnan, P. S.; Richards, J. F.; Skånes, H.; Steffen, W.; Stone, G. D.; Svedin, U.; Veldkamp, T. A.; Vogel, C.; Xu, J. The Causes of Land-Use and Land-Cover Change. Moving Beyond the Myths. **Global Environmental Change**, v. 11, n. 4, p. 261-269, 2001.

Neter, J.; Kutner, M.H.; Nachtsheim, C.J., Wasserman, W. **Applied Linear Statistical Model**. New York: McGraw-Hill, 1996. 1408 p.

Veldkamp, A.; Fresco, L. O. CLUE-CR: an integrated multi-scale model to simulate land use change scenarios in Costa Rica. **Ecological Modelling**, v. 91, n. 1, p. 231-248, 1996.

Veldkamp, A.; Lambin, E. Predicting land-use change. **Agriculture, Ecosystems and Environment**, v. 85, p. 1-6, 2001.