

Göran Djurfeldt, Venkatesh Athreya, N. Jayakumar, Staffan Lindberg, A. Rajagopal, R. Vidyasagar

Modelling social mobility in rural Tamil Nadu

Annexe to the article: ‘Agrarian Change and Social Mobility in Tamil Nadu’ published in Economic and Political Weekly, 2008, Volume 43, No 45. pp. 50 - 61

Abstract

This paper is an annexe to the article ‘Agrarian Change and Social Mobility in Tamil Nadu’ which presents an analysis of the social transformation over a period of 25 years in six villages in the former Tiruchy District in Tamil Nadu. The two most important external drivers are local industrialization and social policy in a broad sense. It is shown that the overall effect is a centripetal tendency in agrarian structure, with tendencies towards a strengthened position for family farming and for the underdogs in the old agrarian society to leave agriculture altogether, seeking improved life chances in the non-agrarian economy, both inside the villages and in the wider economy. In this paper, which is annexed to the article, we present the statistical analyses of mobility matrices by means of regression techniques, which corroborate and elaborate the results and in some respects fail to confirm them.

This is a sequel to the article “Agrarian Change and Social Mobility in Tamil Nadu” (2008), in which we described the changes in agrarian structure in what used to be two Panchayat Unions in the undivided Tiruchirapalli District in Tamil Nadu, India. The first panel wave was in 1979/80 when we interviewed members of a sample of agrarian households in six villages, three in the canal irrigated area along the river Kavery and three villages in the dry hinterlands south of the river. In 2004/05 we again interviewed these households or their descendants with surprisingly low rates of attrition. This means that we have a panel over a generation, which should be eminently fit to analyze changes over the same period.

In Djurfeldt et al. (2008) we describe three trends: (i) local industrialization, (ii) the structural transformation of the rural economy and (iii) the importance of social policy in a wide sense for the outcomes of these processes. We recognized that a descriptive, largely bivariate statistical analysis is rather inadequate for causal analysis. Bivariate association can always imply spurious correlations and hide rather than highlight underlying causal relations. In this paper therefore, we attempt to substantiate the descriptive analysis by means of an exercise in statistical modelling.

We start by giving relevant details of the area, the fieldwork, and the panel.

The area and the fieldwork

In 1979/80, Athreya, Djurfeldt and Lindberg (1990) studied six villages in Tiruchi District, Tamil Nadu and did a detailed survey of, among others, a main sample of 240 households. Twenty-five years later, the current authors returned to the same villages, set on tracing the original participants and their descendants. A pilot study conducted during the summer 2003 and involving a sub-sample of the original households showed that it was indeed possible to trace a surprisingly large proportion of the original sample (well above 90 per cent).

Therefore, attrition was not expected to be an insurmountable problem.

We launched a full-scale resurvey during the autumn of 2005. Treating the 1979/80 study as a base-line, we created a panel database of both qualitative and quantitative data. In the following sections, we will go through the methodology of the panel study, the sampling strategy and the weighting system, attrition and other aspects necessary to judge the validity and reliability of the results reported in later sections.

The original sample is a multi-stage one, beginning with the purposive selection of two units, Manaparei and Kulithalei Panchayat Unions in what was then Tiruchy District. The idea was to select a contiguous and relatively small area containing the variance between dry, rainfed tracts and the “wet”, irrigated areas which are so typical, not only of Tamil Nadu, but of much of South and Central India.

The second stage was the selection of six revenue villages within the first stage units. This was done with so-called PPS sampling, i.e. with *probability proportional to size*.¹ As a result we got three “wet” and three “dry” villages. At the third stage, finally, we selected 40 households in each village with simple random sampling (SRS). Deducting two refusals, we got 238 households. This was the main sample.

Variance estimators are standard formulas in simple random samples, but in multi-stage sampling, the formulae have to be worked out for each specific sample design. As demonstrated in Athreya et al. (1990, p. 47 ff.), we can use standard SRS-estimators, provided we take account of the *design effect* (Kish 1957), which is a function of both the number of stages and the type of sample (PPS, SRS etc.). For variables which are not too skew, Athreya

¹ As discussed by Athreya et al. (1990), there were some impurities in design which, however, were not important enough to imply the more cumbersome way of treating the sample as PPZ, i.e. with units selected with *probability proportional to estimated size* (1990, p. 47 ff.) See Cochran for a treatment of PPS and PPZ sampling (1977).

et al., have shown that the ‘design effect’ in their sample can be taken to be about 1.3. What this means is that, if we want to avoid *false positives* or Type I errors, 30 per cent should be added to a confidence interval and to the critical value of the test statistic. For 2004/05 data, since a further sampling stage is involved, the design effect may be higher.

Since the two ecotypes - wet and dry - are substantially different in their agro-ecological as well as social and cultural characteristics, they should be treated as two different universes for the purposes of arriving at estimates of a range of characteristics. Estimates should therefore ordinarily be made separately for the two ecotypes.

Creating the panel²

Attrition is a problem in all panel studies, since a portion of the original units disappear from the population, either by passing away or by emigrating from the area. Over a whole generation, the problem is likely to be severe. However, we did not want only a panel, but also a new cross-section. This was to enable us to compare the cross-sectional sample made in 1979 with a similar cross-section in 2004/2005, but containing the surviving units and *their descendants*. If there were more than one descendant household, we randomly selected one of them to replace the original one. Moreover, we tried to trace households who had migrated from the villages.

To make the 2004 sample representative of the current agrarian population, we made lists of households who had settled in the village since 1979 and drew a sample of these.

In many settings, the ambition of tracing households after 25 years would have been in vain. However, given the limited geographical mobility and the impossibility of remaining anonymous in a village setting, it proved easier than expected to trace almost all of the original main sample households. Thus we traced 233 households out of the originally interviewed 238 main sample households, the remaining five had become extinct. Of the 233 households traced, some still remain under the same head, which in this context normally is a male, and most of them remain in the agricultural sector. Others remain in the village and the sector, but have a new head. Still others have passed through a generational transfer where landholdings and other property have been partitioned between the heirs, normally among the sons of the former head. Yet 31 households have emigrated, but left enough traces in the village to enable us find out to where.

Of these 233 households, 20 have left agriculture since 1979/80. The remaining 213 make up our sample of resident agricultural households in the study and is the main source of statistical analysis. This sample we call the agrarian population.

A questionnaire was developed, largely following the instrument used in 1979/80. Data collection started in September 2005 and ended in February 2006 but referred to the crop year 2004/05.³

We judge the quality of our data to be high. This is due first of all to the work we put down in 1979/80 when we thoroughly cross-checked all information on important variables, like landownership, with registers, neighbours etc. The enthusiastic response we received when returning in 2004 contributed to data quality.

² See Menard’s short book for a good treatment of methodological issues in panel design (1991, Ch. 3).

³ An excellent job was done by a team of investigators, most of them post graduates in Economics from Bharatidasan University in Tiruchirapalli. Mr. M. Dharmaperumal, Madras Institute of Development Studies, Chennai helped making the data entry formats.

Analytical aspects of mobility studies

The notion of mobility presupposes, not only longitudinal data and a time-line at the end of and during which, the unit studied (the farm or farm labour household in our case) moves upwards and downwards in some well-defined sense. The whole *problematique* has been related to ideological discourses on the development of society. At one end of the spectrum we have the proponents of polarization theories, some of them claiming a Marxist lineage, although some of the Marxist classics were more circumspect in this regard (Djurfeldt 1981).⁴ At the other end, we find theories predicting a centripetal movement in the class structure. Both these approaches have fundamental implications for the reproduction of the social order and both, we contend, are mistaken in positing universal trends or tendencies. Outcomes are conditioned by contextual factors of an economic, political and perhaps even ideological nature. Thus there are good reasons to try to understand not only mobility trends but the various drivers of mobility, a task we now attempt.

Another presumption of mobility studies is a *social space* within which the study unit moves. How is this space to be defined? Classically, mobility studies have defined the social space primarily in terms of class, with widely varying notions of class, including for example educational classes. In methodological terms classes are always discrete and usually hierarchically ordered, which is why statistically we end up with ordinal scales and cross-tabular data.

Such a *discrete and ordinal* conception of social space yields mobility matrices, with two ordinal-scale or binary axes, the x-axis denoting original class and the y-axis denoting current class. Before the development of the personal computer, this approach was more or less a practical necessity, since the resulting data format could be handled with the computing aids then available. With contemporary computing resources, a multivariate extension of the classical approach can be a linear regression, a loglinear or logistic model where the logarithm of the odds of moving from one “class” to another is regressed on a number of independent variables.

Another approach applicable for example to mobility in the landownership hierarchy is movement between *quantiles*, for example deciles or preferably perhaps quartiles. Here loglinear or logistic modelling may also be used, as in the first approach.⁵

By including the lagged value of the dependent variable (y_{t_0}) among the independent variables in a regression, we get an *autoregressive model* (Menard 1995). By means of such a model we can control for autocorrelation⁶ in the dependent variable. While such an approach has many advantages, the drawback is that coefficients of determination get inflated and tests of the regression become biased.⁷ This bias can be checked by running the model with and without the lagged variable. One advantage of autoregressive models is that we can control for historical factors. In other words, what happened before 1979 is controlled for by the lagged variable.

⁴ See also (Patnaik 2007).

⁵ This discussion draws heavily on (Formby, Smith et al. 2004; Goldthorpe and et al. 1980; van de Gaer, Schokkaert et al. 2001).

⁶ Autocorrelation is often a problem in time-dependent data where the values of a variable at one point in time depend on the values of the same variable at a previous point in time.

⁷ The coefficient of determination (R^2) measures the proportion of variance in the dependent variable that can be attributed to the independent variables. Regression routines contain tools to test the zero-hypothesis that the proportion of variance accounted for is merely due to random effects (Anova in OLS and χ^2 in logistic regression).

By including the age of the head of household, we control for cohort effects. If the age factor is significant, it signals that at least part of the change in the dependent variable is associated with life-cycle effects, i.e. of households heads being older today than they were in 1979. Independent variables should either be measured at t_0 and, if they deviate too much from normalcy, they may need to be logged (in the case of ratio or interval variables).

Alternatively, they should capture some other measure of change, like change in the share of labour devoted to non-farm activities or change in the share of income from non-farm sources (since 1979).

As a control variable one can introduce class as operationalised in 1979/80 and see if it contributes to the explanatory power of the above. This would be in the form of a number of dummy variables, with one reference category, for example middle peasants or family farmers. The same approach, i.e. using dummies would apply also to caste, ecotype and other nominal or ordinal scales.

As we will see, the detailed models will be different: For entry into and exit from cultivation (Model 1), we will use a logistic regression model, estimated with Maximum Likelihood.⁸ For modelling movement between different size-classes of holdings (Model 2), we will develop a trichotomous dependent variable making it necessary to use a multinomial logistic regression. Such a model contains no autoregressive component. (More about this below).

Poverty alleviation finally is studied by means of Model 3, where the natural log of total income in 2004 is regressed on a number of independent variables, including an autoregressive component, i.e. the log of total income in 1979. Ordinary Least Squares and bootstrap sampling is used for estimation. Each of these models involve complications that are further discussed below.

As a rule of thumb in regression analysis, it used to be said that a minimum of 10 – 15 cases per independent variable would be necessary to get robust estimates and for testing hypotheses. With around 200 cases, as we have, this means that we could squeeze in 15 – 20 variables for an ordinary regression, which is more than we need. Since in the multinomial model, we get two equations, there are more narrow constraints there, but still cases enough. There is another limitation, however: It is not possible for example to use half the sample to develop a model and then the other half to test it. The only way is the more *ad hoc* method of looking for models with good fit, although this necessarily involves the risk of false positives or negatives (Type I and II errors). For the third model, we have attempted to control for this by means of bootstrap sampling.⁹

A final possibility is obviously Omitted Variable Bias (OVB): Not incorporating important variables into a regression model can invalidate the whole model. We leave it to the reader to judge if there are important such omissions.

Mobility between and within generations

Studies of mobility usually focus on mobility between generations. Although we have two panel waves with 25 years between them, less than half of our households have gone through a generational transfer during the interval. Thus we have 105 households, which remain unpartitioned, in 11 cases under a new head.

⁸ Maximum Likelihood Estimation is more robust and less sensitive to deviations from linearity in the independent variables than Ordinary Least Squares.

⁹ This is a method where a large number of repeated samples, usually of the size $n-1$, are drawn with replacement from the original sample. Confidence intervals can be calculated for bootstrap estimates of regressions coefficients (β -values) to test the zero-hypothesis that they are not significantly different from zero. With the method, the risk for false positives is considerably reduced.

On the other hand, we have 97 households, which are descendants to households interviewed in 1979/80. For descendant households, then, we study inter-generational mobility, for example the difference in landownership between the descendant and the ancestor. For “old” households we study intra-generational mobility. Differences between the two sub-samples can be statistically controlled for in the regression by using a dummy variable for partitioned/unpartitioned status.

Model 1: Entry into and exit from farming

We will proceed by looking at mobility between being a cultivator and not being one. In other words this deals with exit from and entry into farming. In Djurfeldt, 2008 #1714, Table 4} we found a net mobility out of agriculture which was higher for partitioned household than for non-partitioned ones. As we will see, this relation disappears when we control for other factors.

We can try a regression model for the above. We will be looking at the odds of being a cultivator in 2004.¹⁰ We take the logged odds of being a cultivator in 2004 (y_{t_1}) as a function of a vector of independent variables. Thus we will work with the following binary logistic regression model:

$$E(y_{t_1}) = \log_n(odds_q) = \alpha + \lambda y_{t_0} + \beta_1 x_1 + \dots + \beta_n x_n,$$

where:

$E(y_{t_1}) = \log_n(odds_q)$ = the estimated natural log of the odds of being in a given position in 2004, i.e. being a cultivator.

α = constant

λ = regression coefficient for:

y_{t_0} , i.e. the value of the dependent variable in 1979

β_i = regression coefficients for:

x_i = independent variables, either measuring change from 1979 to 2004, or measured in 1979.

Independent variables are:

- the autoregressive component, i.e. having been a cultivator in 1979 (y_{t_0});
- having partitioned since 1979, i.e. a dummy¹¹ (x_1);
- age of head of household, logged (x_2);
- ecotype, a dummy coded 1 for wet and 2 for dry villages (x_3);
- Caste grouping, with dummies for Scheduled Caste and Backward and Most Backward Caste (x_4, x_5);
- Class in 1979, using three dummies: Agricultural labourer and poor peasant (x_6), family farmer (x_7) and big farmer or landlord (x_8). The reference category is “Other and uncodable”;
- The change in proportion on non-farm income from 1979, i.e. a difference between two proportions, ranging from -1 to +1 (x_9);

¹⁰ Odds are defined as the probability of an event, divided by the probability of its non-occurrence, i.e. $p/(1-p)$. Odds move from minima close to zero for unlikely events infinitesimally approaching ∞ for highly likely events. In the equation we work with the logged odds ($\ln(p/(1-p))$), which move from $-\infty$ to $+\infty$. Estimation in this type of models is done by Maximum Likelihood (ML). Unlike Ordinary Least Squares regression ML makes estimates less sensitive to deviations from normality.

¹¹ A dummy is regression terminology for a binary variable, usually coded 0 and 1.

- Joint family, a dummy (x_{10}).

We will be running two models, one without and one with the autoregressive component (y_{t_0}). This enables us to control for the effect of autoregression on the coefficient of determination (R^2) and the associated statistical test (Hosmer Lemeshow test). In the first model we include only ecotype, class and caste as control variables but not the autoregressive component (y_{t_0}). By using this approach we can control for factors which were fixed before 1979, as were ecotype, caste and class. The estimates of such factors are not unbiased in the autoregressive Model 1b, because the residual in y_{t_0} is multicollinear with these factors.

As can be seen from Table 1 below, ecotype comes out as a highly significant factor, with a much higher probability for household to be a cultivator in the dry area – a non-surprising finding given the lower rate of landlessness there. The programme tests the overall statistical significance of the categorical variables, i.e. class and caste, but reports no β -values or antilogs for them. Only class comes out as statistically significant (marked with *** in the table).^{12,13} At the same time we see that the β -values for the three classes (with other and uncodable as reference category) do not attain statistical significance, although they have expected signs, with negative values for both agricultural labourers and poor peasants. Since differences are not statistically significant this does not decisively support our hypothesis about a centripetal tendency in the agrarian structure.

Table 1. Binary logistic regression model for cultivation status in 2004 (Model 1).

	Model 1a			Model 1b		
	β	s.e.	exp(β)	β	s.e.	exp(β)
Cultivated some land in 1979				1.818	0.516 **	6.159
Partitioned holding since 1979				0.345	0.476	1.412
Change in share of non-farm income since 1979				-1.398	0.472 *	0.247
Ecotype	2.090	0.397 ***	8.085	2.387	0.448 ***	10.885
Joint family				0.941	0.568 *	2.563
Age of head of household, logged				-1.015	0.765	0.362
Caste, in total						
Scheduled Caste	0.427	0.722	1.533			
Backward and Most Backward Caste	0.550	0.697	1.733			
Class in 1979, in total			***			*
Agricultural labourer or poor peasant	-1.703	0.900	0.182	-0.964	0.976	0.381
Family farmer	0.129	0.931	1.138	0.139	0.965	1.149
Big farmer or landlord	-1.213	0.994	0.297	-0.834	1.024	0.434
Constant	-1.965	0.571 ***	0.140	0.179	3.254	1.196

Note: No. of cases = 202, per cent missing = 13.3, Nagelkerke's R^2 = .393 and .483 respectively. The p-value for the Hosmer Lemeshow test is .366 and .937 respectively.

Caste can in this case be seen as a proxy for class, but the proxy fares worse than the class indicator that we developed with the 1979 data (Athreya, Djurfeldt et al. 1990, chapter 5). To recapitulate, this classification lays emphasis on whether farmers' production is enough to meet their subsistence needs, or if they produce a surplus big enough to replace their input of family with hired labour. There are unknown sources of error in these estimates, both due to data quality and due to the fact that the method does not capture yearly and seasonal swings in farm economy. The current result that the indicator works better than the proxy is obviously a vindication of our classification. Given this result we exclude caste from Model 1b.

¹² From here onwards we use *** to denote statistical significance at .1% level, ** for 1% and * for 5% level of significance.

¹³ This should be read as a test of the hypothesis that the reference category differs from all others.

Model 1b is an autoregressive model with (y_{i_0}) among the independent variables. All remaining x-variables have been entered, but we have used the standard in the SPSS logistic regression module, which would include a variable only if it passes a threshold test for inclusion and yet another test for removal.¹⁴ As can be seen from the figures given in the note to the table, the inflating effect on the coefficient of determination and the associated test is not very substantial. Comparing with Model 1a we can conclude that the model is statistically significant and that, going by Nagelkerke's R^2 , it accounts for almost 50 per cent of information in the dependent variable. The bias resulting from multicollinearity between the autoregressive component and ecotype and class, shows up in a lower statistical significance for class, but for ecotype the effect if anything goes in the opposite direction, with a somewhat higher β -value in the second model.¹⁵

In Model 1b we first test for autocorrelation with the result that, as can be seen, if you were a cultivator in 1979 you are more than 6 times¹⁶ as likely to be a cultivator in 2004, compared to somebody who was not a cultivator in the first instance, (other variables kept to their means). Thus, and not unexpectedly, there is strong autocorrelation in the model. However, this does not seem to be a cohort effect, because the log of the age of head of household is not statistically significant. Similarly household partition is not statistically significant. This reflects what we saw already in (Djurfeldt, Athreya et al. 2008, Table 3), viz. that mobility rates differ little between partitioned and unpartitioned households.

Family structure may be both a determinant and effect of livelihood strategies. For a family farm, increasing available labour resources by avoiding partition and keeping the family together is a sensible strategy, especially in the absence of mechanization on a Western scale promoting one-man farms (Blekesaune 1996). By such means farmers can avoid depending on hired labour and can compete with big farms dependent on such labour. Having been a joint family in 1979 would be the first choice to capture this effect. However, the 25 year lag appears to be too long here, because we get no effect. By using the current family structure we do get an effect, but we cannot sort out the direction of causality. Is it the result of past successes in farm and non-farm pursuits, which have made it possible to keep the family together, or is the family structure a determinant of the current cultivator status? It is not vital to sort this out, so we are content to note the correlation between the two variables.¹⁷

¹⁴ The first test uses the probability of score statistic for variable entry where we have used the default of 0.05. The larger the specified probability, the easier it is for a variable to enter the model. To control exclusion of variables, the probability of conditional, Wald statistic is used. The default is here 0.1. The larger the specified probability, the easier it is for a variable to remain in the model.

¹⁵ Model 1b gives a list of three outliers and one extreme case. Removing these from the model inflates Nagelkerke's R^2 to .536 (compared to .483). It furthermore strengthens the autocorrelation and makes it statistically significant at .1% level ($\lambda = 2.754$ compared to 1.818). In this model the difference between partitioned and unpartitioned households is reduced ($\beta = .208$ and .345 respectively). The one-sided test for a cohort effect is similarly non-significant, with a negative β -value of -.749, where the hypothesis predicted a positive association. For the other factors the levels of significance remain unaffected with small changes in β -coefficients. This would strengthen the conclusions drawn.

¹⁶ This is read from the value of the antilog, $\text{Exp}(\beta)$ in the last column of the table. Cf. note 19.

¹⁷ We have tested and excluded a few other variables from the models. Firstly, we may have cases where a farming household has divided its holdings among the heirs, but where the father continues to be active in farming his own land. We check for this by testing the effect on cultivator status of cases where the head of household is above 60 years of age. We get no significance. Secondly, we have looked at female-headed households, but find no statistical significance. Finally, we attempted to test if the structural transformation of the economy has accelerated over the 25-year period by looking at possible differences between households established before 1980 and before 1990 respectively. Again we get no significance, which is why we have excluded the variable from the model.

The change in share of non-farm income between 1979 and 2004 would be our attempt to capture the influence of pluriactivity.¹⁸ If our hypothesis is correct, an increase in the share of non-farm income would indicate a household strategy focused on the non-farm sector, and it would thus increase the likelihood of exiting from agriculture and is expected to be negatively correlated with the dependent variable. What is the null hypothesis? It would be that some kind of push factor, e.g. proletarianization or general distress, push people out of farming. Our material does not support this common interpretation. Instead our hypothesis is borne out by a negative β -coefficient, statistically significant at below 5 per cent level. As furthermore can be seen, an increase by one unit in the share of income from non-agricultural sources decreases the odds of currently being a cultivator, more precisely by 75% ($1 - \text{Exp}(\beta)$).¹⁹ This supports the hypothesis that pluriactive households are more likely to move out of cultivation, unlike a generation ago when, we believe, non-agrarian incomes were more likely to be re-invested in agriculture (cf. the discussion of the growing importance of the non-farm sector in Djurfeldt, Athreya et al. 2008). (Besley and Burgess 2000)

In order to test the hypothesis that the underdogs and the topdogs are more likely to move out of farming than others, we again have to look at probabilities conditioned by class. Here the β -coefficient estimates from Model 1a are better than those in the other model. The class factor is statistically significant at 0.1% level, although the β -coefficients for the individual class dummies are not statistically significant. Although, the β -coefficients have the expected signs the hypothesis about topdogs and underdogs tending to leave is not supported by these data.

Model 2: Mobility between size-classes of operational holdings

The second of the three models built in this paper, deals with mobility in operated area. We start with the following size mobility matrix, or more precisely, two matrices, one for each ecotype:

Table 2. Size-mobility matrices for operated area by ecotype, total per cent

Ecotype		Size-class 2004					Total
		0-1	1-2	2-4	4+		
Wet	Size-class 1979	0-1	38,6	2,8	1,4		42,9
		1-2	11,8	9,0	2,5	0,9	24,2
		2-4	12,8	1,5	4,9	0,9	20,1
		4+	7,8	1,7	1,7	1,5	12,8
	Total		71,1	15,1	10,6	3,3	100,0
		0-2	2-4	4-8	8+	Total	
Dry	Size-class 2004	0-2	23,2	8,0	2,4		33,6
		2-4	15,4	5,8	6,7	1,9	29,8
		4-8	8,9	5,1	13,1	1,9	29,1
		8+	1,9	2,3	2,4	0,9	7,5
	Total		49,5	21,2	24,6	4,7	100,0

Note: No. of cases =200, missing 2%

The population we are dealing with here consists of agrarian households in 1979, including landless labourers and the descendants of these household who still remain in the villages and in the agricultural sector. Emigrants are thus not represented and neither are immigrants.

¹⁸ The change in the use of household labour resources is an alternative operationalization which was tested and then deleted.

¹⁹ β -coefficients estimate how much the logged odds changes with a unit change in the respective independent variable. $\text{Exp}(\beta)$ is the antilog of β and estimates the change in the odds associated with a unit change in the independent variable. Negative β -values denote negative associations and give $\text{Exp}(\beta) < 1$.

Since landless labourers are included in the 0-1 category, the totals give a misleading impression of a general downward movement (cf. the Gini indices reported in Djurfeldt, Athreya et al. 2008).

As can be seen we have divided operated area into ecotype-specific size classes. As usual with these matrices, along the diagonal of the table we find the stable households, totally 54.0 per cent in the wet area and 43.0 in the dry villages. Thus, mobility seems to be higher in the dry area.

The upwardly mobile cases are located above the diagonal, comprising 8.5 per cent in the wet area and 20.9 per cent in the dry one – indicating a more than double rate of upward mobility in the dry ecotype, compared to the wet one. Rates of downward mobility differ little between the ecotypes. Thus the lower stability in the dry villages is compensated for by a higher upward mobility, with comparatively more households having increased their holdings over the last 25 years. This may reflect investments in irrigation, and mobility here can go hand in hand with some exit from agriculture as well as consolidation of holdings of those who remain in it.

We can now pose the question of what factors influence the relative risk of being downwardly or upwardly mobile. Formulated this way, the problematic can be addressed with a multinomial logistic regression model.

While logistic regression was originally worked out for binary dependent variables, multinomial logistic regression is a variety adapted to nominal scales. In ordinary logistic regression the dependent variable is the natural logarithm of the odds ($n\log(p/(1-p))$). In multinomial regression y is similarly the probability of one outcome (say downward mobility or p_1) relative to the reference category (i.e. p_2 = the probability of stability, i.e. of remaining in the same size-class of holding). The dependent variable y_1 of outcome 1, then is $n\log(p_1/p_2)$. For the reference category there is an implicit equation with all β -values = 0 and the intercept equal to the log of the overall odds of stability ($p_2/(1-(p_1+p_3))$). Since the latter need not be reported, the output in this case gives two equations, one for downward and the other for upward mobility.

$$\begin{cases} E(y_1) = n \log(p_1/p_2) = \alpha + \beta_i x_i \\ E(y_3) = n \log(p_3/p_2) = \alpha + \beta_i x_i \end{cases}$$

where:

$E(y_1) = n \log(p_1/p_2)$ = estimated natural log of the relative risk of having been downwardly mobile;

$E(y_3) = n \log(p_3/p_2) = \alpha + \beta_i x_i$ = estimated natural log of the relative risk of having been upwardly mobile;

α = constant

β_i = regression coefficients for

x_i = independent variables.

Independent variables are:

x_1 = operated area in 1979;

x_2 = operated area in 1979, squared;

x_3 = age of head of household;

x_4 = age of head of household, squared;

x_5 = household partitioned since 1979;

x_6 = ecotype;

x_7 = joint family 2004;

x_8 = agricultural labourer or poor peasant 1979;

x_9 = family farmer 1979;

x_{10} = big farmer or landlord 1979;

x_{11} = Scheduled Caste;
 x_{12} = Backward or Most Backward Caste

Unlike Model 1, this model contains no autoregressive component. The dependent variable is the change in size-class between the two waves. Therefore it does not suffer from the tendency of autoregressive models to inflate the coefficient of determination and to underestimate variables like ecotype or caste that are constant over the panel waves. Thus we can proceed directly to the estimation (see the table below).

Table 3. Multinomial logistic regression for mobility between size-classes of operated area (Model 2).

Factors	Downward			Upward		
	β	Std. Error	Exp(β)	β	Std. Error	Exp(β)
Constant	-14.286 ***	3.851		-6.370	4.730	
Operated area 1979, acres	1.173 ***	0.262	3.231	-0.341	0.414	1.407
Operated area 1979, squared	-0.042 **	0.015	0.959	-0.012	0.052	1.012
Age of head of household	0.012	0.076	1.012	-0.063	0.136	1.065
Age of head of household, squared	0.000	0.001	1.000	0.001	0.001	0.999
Household partitioned since 1979	1.787	0.554	5.973	-1.512	0.675	0.221
Ecotype	-1.089	0.631	0.337	1.513 *	0.611	4.542
Joint family, 2004	-3.200 ***	0.758	0.041	-0.514	0.602	0.598
Agricultural labourer or poor peasant 1979	-2.749 *	1.013	0.064	-3.462 **	1.181	0.031
Family farmer 1979	-3.455 **	1.067	0.032	-1.581	1.102	0.206
Big farmer or landlord 1979	-4.846 *	1.722	0.008	-1.451	1.772	0.234
Scheduled Caste	-0.006	0.907	0.994	-2.134 *	1.068	0.118
Backward or Most Backward Caste	-0.750	0.855	0.472	-0.286	0.918	0.751

Note: The reference category is stable. No. of cases = 193, missing = 4,5%. Nagelkerke's R² = 0.57.

Here we test the cohort effect by taking age of the head household (x_3), as well as the square age (x_4), the latter in order to control for a curvilinear relation. None of the terms become statistically significant, either in the downward equation or in the upward one. Once again then, we spot no evidence of a cohort effect. In that sense then, the results to be discussed should not be biased.

Looking at the consequences of partition (x_5), we see no significant difference between partitioned and unpartitioned households, in either equation. In this case too we see that differences between the two categories of households is lower than expected, which probably largely depends on what we saw above, viz. that many holdings pass undivided from one generation to the next one.

We have included the square of operated area in the model in order to test for a U-shaped distribution of mobility probabilities. Both operated area (x_1) and its square (x_2) are statistically significant in the downward equation, while they are non-significant in the other one. In the first case, the β -value is positive for operated area and negative for the square. This indicates an inverse U-shaped distribution: As operated area increases, the relative risk for downward mobility increases up to between 5 and 6 acres at which point the negative square term becomes greater than the unsquared one and the risk starts to fall again. Does this not imply a refutation of the underdog-topdog hypothesis? We will return to that question in a while.

There are no statistically significant differences between the ecotypes in the relative risk for downward mobility. In the equation for upward mobility, however, ecotype does come out as significant at the 5% level. This corroborates the bivariate result already reported.

As can be seen from the table, the comparison with the reference category gives a ranking of the three classes in terms of the relative risk for downward mobility, with the lowest risk for big farmers and landlords, an intermediate position for family farmers and the highest relative risk for agricultural labourers and poor peasants. In the upward equation only agricultural labourers and poor peasants face a relative risk of upward mobility lower than that for the reference category.

The results thus show that all classes have lower relative risks of downwardly mobility than the reference category (other and uncodable). However, only the difference between agricultural labourers and poor peasants, on the one hand, and big farmers and landlords is statistically significant. The difference between the latter two classes and the family farmers is not statistically significant. Taken together the two equations imply that agricultural and poor peasants faced the lowest relative chances of upward mobility and the highest relative risks of losing position during the period 1979 to 2004. The higher classes faced lower relative risks. We spot no differences between family farmers and others in this regard. This makes intuitive sense in that households with better asset positions, other thing being equal, tend to fare better over time. Policy intended to help the poorer asset categories can be consistent with strengthening of the higher asset categories.

Do these results not contradict our hypothesis about the topdogs leaving? Partly yes: This model deals with the mobility chances of those remaining in agriculture, while Model 1 deals with entry into and exit from the sector. Both models fail to support the topdog-underdog hypothesis. For households remaining in the sector, the dice is weighted against the underdogs, while the topdogs have fared better. The evidence is inconclusive for the family farmers.

Note however the trap which lies in assuming that the classification made in 1979 is a constant. If the classification exercise were to be repeated with 2004 data, some of the 1979 big farmers might have ended up as family farmers, which would have led to other conclusions. We hope to return to this issue in a future paper.

We glimpse a similar pattern for caste, where upper castes are the reference category, but here there is no statistical significance, except for Scheduled Castes and their chances of upward mobility, which are lower than for other castes. Thus, discrimination against the ex-Untouchables remain in force inside the agrarian sector.

Besides class and caste, what are the determinants of mobility? What about non-farm income or labour resources spent in non-farm employment? Furthermore, what about farm investment, e.g. in irrigation equipment? We have tested three indicators of these drivers and none turned out significant. In other words, we cannot show that an increase in the share of non-farm income has any influence on mobility. Using instead the change in the proportion of family labour resources devoted to non-farm activities as an indicator likewise throws up no significant results.

Taking change in the number of pumpsets owned as a proxy for farm investment in general similarly fails to show an significant relation to mobility chances. However, when we remove 9 outliers from the model, we get the expected result: A unit change in the number of pumpsets owned, increases the relative risk (chance) of upward mobility more than 4 times.²⁰ As can be seen from the table, the hypothesis that family type is related to relative risks of mobility gains support from our data. Joint families run a significantly lower relative risk of having been downwardly mobile! How are we to interpret this finding? Joint families obviously have more plentiful labour resources than others and we would argue that this is the decisive factor. A larger pool of labour and non-labour resources permits diversification and hence lowering of risk. Also important, however, is demography in the sense of ratio of dependents to earners.

The above is in line with our general hypotheses that local industrialization implies increased competition for labour between agriculture and industry. Similarly, the whole range of social policy interventions decreases the risk for poor people to land in client relations with the local

²⁰ Removing these outliers from the model resulted in no other substantial changes. This indicates that the model is quite stable and, above all, that the unexpected results for family farmers is not due to weaknesses in the model estimated.

rich. The fact that permanent farm servants (*panneial* in Tamil) have gone down drastically in numbers is a clear testimony to this development. Managing a farm business is much facilitated by access to enough family labour and gives a competitive edge to joint families. A lower risk of downward mobility of joint families could have been an indicator of *family farmerization*.²¹ However, the fact that we have been unable to show that family farmers suffer less risks of downward or upward mobility than others implies otherwise: All classes may enjoy the economic advantages of a joint family. More specifically we cannot show any difference between family farmers and big farmers and landlords in this regard.²² Moreover and as already pointed out, big farmers and landlords may also have been “family farmerized” since 1979.

To conclude, this model qualifies somewhat the hypotheses we are driving in this paper. The first model confirms the effect of the structural transformation of the economy but gives no conclusive support to the topdog-underdog hypothesis. Model 2 on the other hand shows that for those remaining inside the sector, mobility chances are still in favour of the big farmers and against the agricultural labourers, poor peasants and the Scheduled Castes. Only by leaving the sector, can the latter escape from the discrimination against them. We have not been able to demonstrate that the advantages enjoyed by family farmers have implied greater chances of upward mobility.

Model 3: Drivers of poverty alleviation

As our article shows (Djurfeldt, Athreya et al. 2008), real incomes for agrarian households in 2004 have improved a great deal compared to 1979 and more so for farmers than for agricultural labour households. At the same time inequality seems to have gone down. By implication, poverty should also have gone down. What are the drivers of this development? In line with our hypotheses, we are interested in separating the effects of (i) local industrialization and the structural transformation of the rural economy (ii) agricultural growth and (iii) social policy. We would like to see if some of the effects of targeted schemes, like pensions for widows and agricultural labourers can be traced in the autoregressive model presented below. However, in so far as social policies take the form of general rather than targeted interventions we cannot trace their effects by means of our material.

We started working with a logistic regression model where the dependent variable was a poverty dummy for 2004 and with poverty status in 1979 as an independent variable, both based on official definitions of poverty. We had to give up this design, because the resulting models were not stable. A scale dependent variable, viz. total income, proved to work better than the poverty dummy. When we take the natural log of this variable, the resulting model gives an almost normally distributed residual, not-too-high levels of multicollinearity and no outliers. Moreover, β -values tend to be stable, or vary in predictable manners between different model specifications. To bring down the likelihood of false positives or negatives, we tested the model by means of bootstrap sampling, drawing 1000 samples from the original, with replacement and with size $n-1$. In the table below we report the mean β -coefficients and their statistical significance in the bootstrap sample.²³

²¹ Cf. Djurfeldt (Agresti 1996, ch. 5) who has shown with Swedish data that farm family structures tend to be adapted to the needs of the farm, also in settings dominated by family farms and by nuclear family ideals.

²² We have tested a model where we included the interaction effects between class and joint family, but we did not get any statistically significant results.

²³ Bootstrap sampling is not applicable to the logistic models earlier used. On the other hand, logistic models based on Maximum Likelihood estimation are less prone to estimation errors arising from non-normality of the variables implied (Besley, Pande et al. 2007).

Thus the model below has the following mathematical form:

$$E(y_{t_1}) = \ln(y_{t_1}) = \alpha + \lambda \ln(y_{t_0}) + \beta_i x_i,$$

where:

$\ln(y_{t_1})$ = total income in 2004, logged;

α = constant;

λ = the regression coefficient for the autoregressive variable;

y_{t_0} = total income in 1979, logged

β_i = regression coefficients for other independent variables;

x_i = other independent variables.

The independent variables are:

x_1 = age of head of household 2004, logged;

x_2 = household partitioned since 1979, dummy;

x_3 = size of household 2004;

x_4 = household exited agriculture since 1979, dummy;

x_5 = household upwardly mobile in size-class of operated area since 1979, dummy;

x_6 = ecotype, dummy (0 = wet village, 1 = dry village);

x_7 = Scheduled Caste, dummy;

x_8 = change in share of non-farm income since 1979;

x_9 = change in number of non-farm workers in household since 1979;

x_{10} = literacy dummy for head of household 1979;

x_{11} = incomplete nuclear family 2004, dummy;

x_{12} = professional agricultural labourer 2004, dummy;

Note first that the income variables are skewed to the right, implying problems with heteroskedasticity.²⁴ Thus a logarithmic transformation is called for so that logged income in 2004 is the dependent variable, and logged income in 1979 is the autoregressive component. With the dependent variable logged, it is preferable to log the independent variables as well. However, this does not work for dummies and for scale variables with negative values. Thus we can only log one of our independent variables, viz. age.²⁵ This means that the other β -coefficients will catch the impact of a unit change in an independent variable on the log of income in 2004.

Model 3 like the first model in this paper is an autoregressive one. To repeat, autocorrelation tends to increase the coefficient of determination and the associated test. Similarly it tends to underestimate the influence of constant factors, such as ecotype or caste. Therefore we started by estimating a model resembling Model 1a, i.e. without the autoregressive component but including ecotype, caste and class. Thus, as long as we do not control for autoregression, we cannot spot any significant differences between ecotypes, between caste groupings and between classes. This is acceptable given that, as we have already seen, incomes have risen faster in the dry area, largely eliminating the economic differentiation between the ecotypes that previously was so stark. Similarly, it can be shown that variance in income is much higher within caste groupings than between them.²⁶ Finally, given the mobility in class

²⁴ Regressing two right-skew variables often yields a fan-shaped scattergram, implying that the variance around the regression line is *heteroskedastic* and higher at the upper end of the scale. This again gives lower precision to estimates and makes it more difficult to find significant β -coefficients for such variables.

²⁵ Actually, household size should have but has not been logged. This implies an underestimated β -coefficient for this factor.

²⁶ Data withheld for reasons of space.

structures, it should perhaps not be expected that class position in 1979 would have much of an influence on income in 2004. Since this model threw up no significant factors for the variables mentioned, it is not included below.

Instead we reproduce only one model, including the autoregressive component and a number of independent variables.

Table 4. Multiple regression model for total income in 2004 (Model 3b), bootstrap means²⁷

Variables	Mean β , weighted
Constant	7.616 ***
Age of head of household 2004, logged	0.447
Household partitioned since 1979, dummy	0.104
Size of household, 2004	0.126 ***
Household exited agriculture since 1979, dummy	0.264
Household upwardly mobile in size-class of operated area, dummy	0.566 ***
Village ecotype	-0.088
Scheduled Caste dummy	-0.221
Change in share of non-farm income, since 1979	-0.218
Change in number of non-farm workers in household, since 1979	0.107 *
Literacy dummy for head of household, 1979	0.039
Incomplete nuclear family, 2004	-0.342
Professional agricultural labourer, 2004	-0.644 ***
Total household income 1979, logged	0.053

Note: No. of cases = 157, missing cases = 22%, adjusted $R^2 = 0.490$.

Given the results of Model 3a already discussed, it is perhaps not so strange that in the autoregressive model we do not get any statistical significance for the autoregressive component (y_{t_0}). In addition to the factors mentioned this is probably a reflection of measurement error and of the fluidity in the structure of incomes. Similarly, we do not get statistical significance for age (x_1), which is why any error due to a cohort effect can be left aside.

As in the other models, we do not get statistical significance for household partition (x_2), which supports the conclusion already reached, viz. that generational transfers are handled so that they do not on the whole increase poverty risks for the descendants.

Furthermore, we have included the dependent variables in Model 1 and 2 as independent variables in this one. This allows us to test hypotheses about the effect of entry into or exit from farming and of mobility in size-classes of operated area on real incomes.

Taking the exit variable (x_4) first, it has to be kept in mind that out of all households who have exited cultivation since 1979, a significant proportion have left the village and/or the sector. Only the remaining households are part of our sample and thus implied by the variable. Under these circumstances exiting would imply either proletarianization (i.e. exiting farming and selling or renting out the land) or becoming a landlord (i.e. exiting farming and leasing out the land). In the former case, if exiting is not associated with increasing income from the non-farm sector, one would expect a decrease in real income. In the latter case, there may be both winners and losers. Thus it is difficult to specify a definite hypothesis to test with the exit variable.

Looking at the result, the exit variable does not reach statistical significance, thus indicating that the households who have not left the sector and/or the village have on the average neither gained nor lost from the transition.

²⁷ Change in the number of non-farm workers is significant for a one-sided hypothesis only.

The dependent variable from Model 2 (x_5) becomes a very powerful independent variable in the current model. A household that has been upwardly mobile in size-class of operated area is likely to have gained a lot in terms of income; in fact, the β -coefficient points to a more than 70 per cent increase in income. What lies under this result?

As we have seen above there seems to have been a more or less steady growth in agriculture over the 25 year period, at the same time as the inequality of the distribution of farm incomes seems lower today than it was 25 years ago. A farm investment proxy, change in the number of pumpsets owned does not turn out significant when upward mobility is included in model, but it does so when the mobility indicator is removed. This again would indicate that agricultural investment and growth have been significant for poverty reduction.

To this interpretation could be added the fact that high-yielding varieties of paddy are now universal and benefiting smallholders as well as others, in line with Lipton's and Longhurst's findings (2005; 1989). However, looking at those farmers who were growing traditional varieties of paddy in 1979 and thus adopted high-yielding varieties since then, we get no statistically significant results, even when the mobility indicator is removed. We recall the results of classical diffusion studies, saying that late adopters of an innovation typically become adopters not to gain from it, but in order to avoid losing (Rogers 1983). This result may be an indication of such an outcome. If anything and more generally, these results would indicate that the Green Revolution has lost its poverty profile and that more capital-intensive growth patterns are now prevailing, especially driven by investments in irrigation. Obviously they have contributed much to growth in real incomes.

Looking now at the effects of the transformation of the economy, we first see that the change in the share of non-farm income (x_8) gives a negative but non-significant β -coefficient. This is an unexpected result and we will not try to find an *ad hoc* explanation for it. On the other hand, we get the expected sign and significance at 5% level for the other indicator, i.e. change in the number of non-farm workers in the household (x_9). The β -coefficient indicates that an increase by one non-farm worker is associated with about 10 per cent increase²⁸ in total income. Weighing the evidence, we think these two tests gives additional although not very strong support to our hypothesis that the real income effects of the structural transformation of the rural economy are positive, rather than negative as expected from the widespread theories of pauperization and distress migration.

As already mentioned, ecotype, caste and class were not statistically significant in a model without the autoregressive component. We have kept ecotype (x_6), and the Scheduled Caste dummy (x_7) in the model reproduced above. With the logic of an autoregressive model, the autoregressive component (y_{t_0}) controls for any influence of ecotype and caste before 1979

(t_0). A statistically significant result would indicate that the factor concerned has an association with income changes since 1979. As can be seen from the table, ecotype is not significant and neither is Scheduled Caste, when the design effect is accounted for. How is this to be interpreted?

Taking Scheduled Caste first, a possible interpretation of a non-significant β -coefficient may be that, thanks to social policy interventions, being a *dalit* is no longer as big a handicap in economic terms. It means that the main drivers of poverty-alleviation to a large extent are "caste-blind". In the new non-agrarian economy, caste discrimination is much less than in the old agrarian society.²⁹ Similarly, social policy interventions targeted towards the poor would, if these results have a more general bearing, neither discriminate against nor be affirmative

²⁸ Taking the antilog of the β -coefficient of .107.

²⁹ In fact our data show that caste explained very little of the variance in incomes already in 1979. There was more variance within castes than between them.

towards the Scheduled Castes.³⁰ However, here the β -coefficient is at least on the border of being significant, which would indicate that caste discrimination is not entirely absent, resulting in marginally lower gains from the structural transformation of the rural economy for the SC households remaining in the agrarian population and the village.

The latter interpretation gets corroborated when looking at the professional agricultural labourers (x_{12}), i.e. labourers with no or almost no income earned outside the farm sector. Such labourers are mainly found in the dry villages and they are very often Scheduled Caste.³¹ As the table shows, they have suffered significantly decreased incomes in real terms since 1979. An implication is of course that social policy interventions still have failed to improve life chances for all underdogs, even if they have meant a lot to many.

Ecotype is non-significant, meaning that when other factors are statistically controlled for, the remaining difference between ecotypes are likely to be caused by random factors.

Surprisingly we get no results for education. We have tried both female³² and male education and get significant results for neither of them. One would expect literacy to be a driver for improved life chances but our results surprisingly do not support that hypothesis. It can possibly be because those who have profited from their education have left the village and thus exited the sample.

A final variable to discuss is the importance of the family. Note first that we get high statistical significance for family size, mirroring earlier results about the importance of joint families.³³ All our results, then, point to the importance of command over labour resources for mobility chances. Furthermore we look at incomplete nuclear families (x_{11}), which often point to women-headed households, often elderly widows and occasional widowers and non-married men and women. Again there are social policy interventions targeted to the first-mentioned sub-categories, but the policy impact is at best patchy, as is indicated by the negative, although non-significant β -value associated with this variable.

³⁰ For a different result, see Besley and Burgess (2000)

³¹ A collinearity diagnostics shows that this variable is mildly collinear with incomplete nuclear family. Thus there seems to be a number of non-married in the category.

³² Both education of wife of head of household in 1979 and in 2004.

³³ This variable yields no statistically significant results in this model. (Besley and Burgess 2000; Besley, Pande et al. 2007)

References

- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York, John Wiley & Sons, Inc.
- Athreya, V. B., G. Djurfeldt and L. Staffan (1990). *Barriers Broken: Production Relations and Agrarian Change in Tamil Nadu*. New Delhi, Sage.
- Besley, T. and R. Burgess (2000). 'Land reform, poverty reduction, and growth: Evidence from India.' *Quarterly Journal of Economics* 115(2): 389-430.
- Besley, T., R. Pande and V. Rao (2007). 'Political Economy of Panchayats in South India.' *Economic and Political Weekly*: 661-666.
- Blekesaune, A. (1996). *Family Farming in Norway: An analysis of structural changes within farm households between 1975 and 1990*. Dr.polit. thesis, PC, Department of Sociology and Political Science, University of Trondheim.
- Cochran, W. G. (1977). *Sampling Techniques*. New York, John Wiley.
- Djurfeldt, G. (1981). 'What Happened to the Agrarian Bourgeoisie and the Rural Proletariat Under Monopoly Capitalism?' *Acta Sociologica* 24(3): 167-191.
- Djurfeldt, G., V. B. Athreya, et al. (2008). *Agrarian Change and Social Mobility in Tamil Nadu*. Lund, Department of Sociology.
- Formby, J., W. Smith and B. Zheng (2004). 'Mobility measurement, transition matrices and statistical inference.' *Journal of Econometrics* 120(1): 181-205.
- Goldthorpe, J. H. and et al. (1980). *Social Mobility and Class Structure in Modern Britain*. Oxford:, Clarendon Press.
- Kish, L. (1957). 'Confidence intervals for clustered samples.' *American Sociological Review* 22: 154-66.
- Lipton, M. (2005). *The Family Farm in a Globalizing World: The Role of Crop Science in Alleviating Poverty*. Washington, D.C., IFPRI.
- Lipton, M. and R. Longhurst (1989). *New seeds and poor people*. Baltimore, Johns Hopkins University Press.
- Menard, S. (1991). *Longitudinal Research*. Newbury Park, Calif., Sage Publications Inc.
- Menard, S. (1995). *Applied logistic regression*. Thousand Oaks, Calif., Sage.
- Patnaik, U. (2007). *Marx and his successors on the Agrarian Question*. New Delhi, LeftWord.
- Rogers, E. M. (1983). *Diffusion of Innovations*. New York, Free Press.
- van de Gaer, D., E. Schokkaert and M. Martinez (2001). 'Three meanings of intergenerational mobility.' *Economica* 68: 519-537.