5 Data

The data used is a panel data set by survey design. The next section shortly explains features of panel data and highlights requirements when working with them.

5.1 Working with panel data

For econometric analyses quantitative micro- and/ or macro-level data are needed. Three types of data can be differentiated and used: time series, cross-sectional and panel data. Time series data reflect the measurement of one or more variables at different points in time over the same unit(s) of observation, e.g. individuals, households or firms. In contrast collecting cross-section data means that the same variables are asked from different observation units at one point in time, which of course in practical field work can be a time span like a week during which households in different villages are interviewed. As a kind of combination of these two methods, the collection of panel data is motivated by raising information from the same observation units at different points in time as done in the surveys of the Research Center for Rural Economy (RCRE). The data set received from the RCRE and used in this study will be explained in more detail in chapter 5.2.

Baltagi (2008) depicts in a clear overview the pros and cons of using panel data. Relevant for the study at hand could be the influence of individual heterogeneity, where in this case ‘individual’ could refer to land as the main input factor for agricultural production in rural Hebei.

A clear advantage of using panel data for this study is that they offer more degrees of freedom in order to deal with endogeneity between off-farm employment and on-farm investment and production decisions (Ahituv and Kimhi, 2006; Phimister and Roberts, 2006; Weiss, 1999). Besides offering more possibilities to detect unobserved heterogeneity, panel data also provide more information and variability (in information) than e.g. cross-section data. Normally there is also less colinearity found between variables. Panel data also provide information on adjustment dynamics of the observed units so it is possible to assess the adjustment of household behavior over time, e.g. with respect to macroeconomic and economic policy changes. This kind of data also allows us to draw conclusions about dynamics at the individual level (Deaton, 1997) and to augment across-household information with variation within households (Bowlus and Sicular, 2003). As Shi et al. (2007) point out, panel data sets allow
for the proper estimation of life cycle effects that influence migration and off-farm activities which are both individual based decisions.

Baltagi (2008) suggests that random effect models are well suited for household panel studies where N observation units have been drawn from a large population. As long as the panel was drawn randomly, this type of models allows estimating individual effects that are characterized as random and inference pertains to the population. If omitted variables seem to exist, panel data can be used to obtain consistent estimators (Wooldridge, 2002). Following Deaton (1997) the “quality of land” is in most cases an omitted variable, e.g. if the data collection is not combined with an analysis of soil samples or plot level climatic data.

Following Wooldridge (2002) unobserved effects on the individual level could be the cognitive ability or personal motivation. Also personal preferences or their change can usually not be directly observed. If we think of a household level model, such unobserved effects can be trust or mistrust and altruistic behavior between household members or between neighboring households. Sometimes, such unobserved effects can also exists on the village level as shown by Piotrowski (2009). The author conducted interviews in Quzhou County where village heads revealed, that they mistrust neighboring villages and therefore do not exchange land with them.

As Corsi and Findeis (2000) emphasize, panel data are also very useful to distinguish between the reasons of state dependence.

5.2 The household data set

The data set available and used for the analysis is part of a large comprehensive study conducted by the Research Center for Rural Economy (RCRE) since 1986. The full RCRE sample for 31 provinces and administrative regions in China covers 300 villages and over 22,000 households (Duclos et al., 2008). For the study at hand a data set for Hebei province is used. The annual household survey includes only rural households. Because the variables contain information about households and individuals, this data set is a micro panel data set but the survey follows only households and not individuals over time. Responding households have to keep a daily dairy of all activities, e.g. farming or other household production, as well as income and expenditures. The information is collected once a month by an administrator that is resident to this region and living at the town or township that is the county seat (Duclos et al., 2008). In addition to the household level data, information about the village is collected every year but for the present work only the household level data set is available. It covers the period...
1986 to 2006. The survey was not conducted in 1992 and 1994 due to financial reasons. Consequently 19 years with observations are available to construct a panel data set. Depending on the village size the sample covering Hebei province contains between 100 and 120 households randomly chosen per village. There are 4 different versions of household and village survey questionnaires – one for the years 1986 to 1991, one for 1993, one for 1995 through 2002 and a fourth one which is in use since 2003. Since 2003 the data set also contains information, e.g. regarding education and labor occupations, about every member living in the household.

The number of villages in the mentioned data set for Hebei is 6, for 1987 to 2003, and 11, for 1986 and the years from 2004 onwards. In the first year there are 1100 households in the data set. For the later years this changes to between 600 (1988) and 1091 households (2005 and 2006).

The household sampling procedure is described by Benjamin et al. (2005). On province level the RCRE selected counties from the lower, middle and upper income tercile. Villages are chosen from those county groups based on geographic criteria (village in plain, hilly or mountainous area), location (rural, suburban or urban), and according to main economic activities such as fishery or forestry (Glauben et al., 2008). So only the households interviewed in the chosen villages are selected randomly.

To derive with appropriate statistical and econometric conclusions samples have to be drawn as representative as possible. However, the data set has some shortcomings that need to be pointed out because they influence the selection of analysis methods and the quality of the results.

The data set lacks a detailed household roster (Duclos et al., 2008) and no household sampling weights are available. According to Chen et al. (2004) the sample shows slight attrition over time. But Benjamin et al. (2007) state, that they observed some attrition especially after the years when the survey was not conducted. Yet, it can be assumed that households lost through attrition have been replaced by new households based on random sampling (Benjamin et al., 2007).

Since the survey round conducted in 1993 the RCRE data set contains a variable classifying the household based on generations. This multivariate variable also allows to identify extended families (including relatives) or broken families (e.g. due to divorce). Sometimes the size of households, which remained in the sample for several survey rounds, drops sharply from one year to the next and on first sight there seems to be no reason for this pattern. But one has to have in mind that in China sons leave their parents’ household once they get married (Jia, 2010). If this happens (and this can not be identified for the years before 1993), the sons (and their new household) are no longer included in the sample,
which might cause some loss of information, e.g. on the composition of the original households’ labor force and income. Based on the household IDs 537 households have been identified to stay in the sample over all survey years. This is a valid statement if the household ID is really a unique identifier.

However, it is possible that the ID of a household which dropped from the survey was used for a household that was included as a replacement (see Chen et al., 2003). Using the information from the variable “was this household surveyed last year?” and comparing the number of permanent residents and the age of main laborers between years were the household stated being surveyed a year before and those years were the household states not being surveyed in the previous year up to 27 IDs are identified, which might suffer from the described problem that the same ID is assigned to different cases. Without access to the original questionnaires, it is not possible to further cross check if this was a mistake in data processing or if the cases are really different but the same IDs are used. According to Jia (2010) the RCRE did not document cases of using old IDs for new observation units. An explanation might be that the RCRE was not aware of the difficulties that evolve for long term (panel) assessments from using the same IDs. The RCRE might have put a greater emphasis on replacing households that dropped out of the sample by “comparable” ones to keep the sample as such representative instead of tracking the same observation units over time. Such inaccurateness occurs in statistical data sets and has to be controlled for, if possible, but by and large official Chinese statistics seem not to suffer from falsification due to political influence on statisticians (Chow, 2006).

In future rounds of the survey there might also be the problem that areas which have been classified as rural will be defined as sub-urban or urban due to urbanization and infrastructure development, administrative mergers of villages into townships or towns or the new classification of towns according to population size (Yang and Zhou, 1999). This would also affect the possibility to track households and individuals if villages were dropped from the sample.

No refusal rates are reported for the RCRE data (Benjamin et al., 2005). Benjamin et al. (2007) raise the doubt that some households of the low end of the income distribution are excluded. They argue that this could be the case because the survey protocol requires keeping a diary of e.g. income and expenditures and this could lead to excluding illiterate (low income) households. Also the survey procedure is putting high costs on rich households. In general households with relatively high opportunity costs of time can be expected to be more likely to refuse (Benjamin et al., 2005). As further potential pitfalls of the design of the RCRE survey Benjamin et al. (2005) mention the possibility of inaccurateness in asking for income from household-run business and problems in distinguishing variable and fixed (production) costs during the enumeration.
ever, for the data set available it is not possible to check, if it is influenced by such tendencies because of the secondary nature. Since only the data set as such is available and no information about how the data have been collected it has to be assumed that the data were collected following a statistically sound sampling design.

For the RCRE surveys since 1993 farm households are asked to specify their time allocation. So they can state that they are either: full-time agriculture, mainly agriculture, mainly non-agriculture, or full-time non-agriculture households or allocate their time in another way. For the calculation of poverty incidence and differences in poverty inequality, as done in chapter 6, this self-classification variable is used to separate the observations in two sub samples: full-time farm households (full-time and mainly agriculture) and part-time (mainly non-agriculture and full-time non-agriculture).

Detailed information about all working household member’s time allocation to farming and non-farming activities would allow a clearer differentiation between the degrees to which the household provides labor hours to farming activities. However such data about the individual allocation of working time of farm household members are only available from 2003 onward.

5.3 Summary

Panel data sets provide the possibility to track individuals and their behavior and decisions over relatively long time spans. This offers the chance to distinguish between time-variant and time-invariant individual effects such as individual preference sets or unobservable land characteristics.

By survey design the available RCRE data set for Hebei province is a micro panel data set where households rather than individuals are followed over time. The information provided is used in a cross-section manner for some part of the analyses and its panel data characteristics are utilized in other parts. The data set is representative for the rural areas of Hebei between 1986 and 2006 but due to changes in the variables contents and due to problems in uniquely identifying observation units only those parts of the observations will be used for the empirical assessments in chapter 6 and chapter 7 which can be correctly identified.