


# Yield Trend Estimation in the Presence of Farm Heterogeneity and Non-linear Technological Change

**Journal Article****Author(s):**

Conradt, Sarah; Bokusheva, Raushan; [Finger, Robert](#) ; Kussaiynov, Talgat

**Publication date:**

2014-05

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000092015>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

**Originally published in:**

Quarterly Journal of International Agriculture 53(2)

# **Yield Trend Estimation in the Presence of Farm Heterogeneity and Non-linear Technological Change**

**Sarah Conradt and Raushan Bokusheva**

ETH Zurich, Switzerland

**Robert Finger**

Wageningen University, The Netherlands, and University of Bonn, Germany

**Talgat Kussainov**

Kazakh Agrotechnical S. Seifullin University, Astana, Kazakhstan

## **Abstract**

An adequate representation of the technological trend component of yield time series is of crucial importance for the successful design of risk management instruments. However, for many transition and developing countries, the estimation of the technological trend is complicated by the joint occurrence of three phenomena: (i) a high level of heterogeneity among different farms in a region; (ii) non-linear development of technological change; and (iii) high yield variations as a consequence of high exposure of rainfed agriculture to extreme weather events. Under these situations, the usually applied approach to detrend crop yield data using Ordinary Least Squares is known to be biased. Based on a unique data set of 47 farm yield data from northern Kazakhstan, we evaluated different alternative approaches. First, we consider the use of the MM-estimator, a robust regression technique for detrending. Second, we evaluate the effect of adding information on extreme climate events as an additional regressor. Finally, we consider combinations of the two former approaches and compare the implications of the different aggregation level on trend estimations. The results reveal the importance of using single farm yield data for detrending, because technical trends in Kazakh wheat yields are highly farm-specific. Furthermore, our analysis shows that the estimation of technological trends can be improved by incorporating weather information in the regression model if time series of crop yield data contain severe fluctuations due to occurrence of climatic extreme events. Thus, the presented analysis contributes to an improved crop yield analysis for many developing and transition countries facing similar conditions.

**Keywords:** detrending yield data, MM estimator, weather information, Kazakhstan

**JEL:** G22, C01, Q14

## 1 Introduction

Multiple risks are present in agricultural production and thus numerous non-formal and formal risk management strategies exist to deal with them. Major events such as severe weather shocks require formal approaches to transfer risks out of local communities, regions or even a country. The measurement of production risks and the development of such risk management instruments strongly rely on the availability and quality of historical data. Besides the volatility caused by adverse weather events and other risks, agricultural yields are strongly driven by technological adjustments. This additional source of observed changes and variability in crop yields explained by technological change has to be removed from the time series to quantify risk exposure. Thus historical yield data is “detrended” for many actuarial applications (SKEES et al., 1997). In most countries, fast diffusion of innovations leads to a rather homogeneous pattern of technological change across farms. Thus, many empirical studies use higher aggregation levels to represent technological trends in farm level yields, assuming that the trend on the aggregated level resembles the trend of a single farm (ATWOOD et al., 2003). Such a procedure allows to reduce the scope of the outliers’ problem and represents a pragmatic approach where long-term individual farm data does not exist (OZAKI and SILVA, 2009). At these aggregated levels, crop yield development exhibits a linear trend over time for the majority of empirical problems (for a review, see FINGER, 2010a; HAFNER, 2003, and TANNURA et al., 2008).

However, for many transition and developing countries, the estimation of a technological trend is complicated by the joint occurrence of three phenomena: (i) a high level of heterogeneity among the different farms in a region; (ii) non-linear development of technological change; and (iii) high yield variations as a consequence of high exposure of rainfed agriculture to extreme weather events. Under these situations, the usually applied approach to detrend crop yield data based on Ordinary Least Squares (OLS) is known to be biased (DOUGHERTY, 2011). Differences in accessibility to factor and credit markets, educational level, and managerial abilities contribute to rather slow and uneven technology diffusion. This in turn causes a rather high variation in farm productivity and a heterogeneous technological trend across farms. Moreover, farms in transition countries have been subjected to several restructuring and privatisation rounds, which seriously affected their investment behaviour and led to alternating periods of investments and dis-investments (BOKUSHEVA et al., 2009). Accordingly, a linear trend might not be sufficient to represent technological trends, and the use of higher level, e.g. quadratic or cubic polynomial functions might be required. However, determining an adequate polynomial degree is often ambiguous. In addition, the often harsh climatic conditions and the rainfed agriculture (LEBLOIS and QUIRION, 2013) increases the dependence of the farms’ production output on weather conditions, which in turn leads to many extreme yield observations, thereby complicating

the selection of outliers. Although challenging under such situations, a careful trend analysis is essential because neglecting or misspecifications of trends lead to biased estimations and impede the development of meaningful risk management instruments (see also BOKUSHEVA and BREUSTEDT, 2012).

All these characteristics are relevant for our case study Kazakhstan. A harsh climate, with regular droughts leads to extensive yield losses and a high year-to-year yield variability representing a non-diversified systemic risk. Thus, production risks from the rainfed spring wheat regions of northern Kazakhstan underlying our empirical analysis are particularly distinct.

In this paper we investigate the performance of various approaches to crop yield data detrending for selecting an adequate trend model and obtaining consistent trend parameter estimates under (i) a high level of heterogeneity among different farms in a region; (ii) non-linear development of technological change; and (iii) high yield variations as a consequence of high exposure of rainfed agriculture to extreme weather events. We combine three elements in our analysis in order to improve the specification of technological trends in crop yield data in this context: first, we test the applicability of aggregated yields to capture yield variation caused by technological change. Second, we employ MM – a robust regression technique used in addition to OLS. Third, we investigate how using additional information in the form of an index representing weather conditions might improve a model's predictive power with respect to yield trend identification. Finally, we analyse whether this additional information is also required for trend estimations using the robust MM estimator.

The remainder of this paper is structured as follows. Section 2 provides an overview of the literature. Section 3 describes the data and methods used in the analysis. Our research findings are presented in section 4. In the last section we discuss the results and draw conclusions.

## 2 Literature Review

An important finding of earlier research is that farm yield data are more volatile than county or even national yield data (MARRA and SCHURLE, 1994; FINGER, 2012; COOPER et al., 2009). Idiosyncratic events such as a local disease affect only a small area and are thus only observable in the individual farm yield data. This “noise” may interfere with the underlying trend pattern and impede a consistent technological trend estimation. Aggregation smoothes the yield pattern and thus only systemic events are reflected. Farm-specific variation is “averaged” across a region or a whole country and consequently with an increasing number of aggregated farms the individual impact of a farm is reduced. This effect already takes place while aggregating very small units,

i.e., the yield of a single field. MARRA and SCHURLE (1994) find a reinforcing relationship between size of the aggregated unit and farm-level yield risk. Hence, increasing the aggregation level (i.e. enlarging acreage) leads to a decreasing rate of yield variability. Consequently, a stable estimation of the prevailing trend pattern may be obtained by using higher aggregation levels. Thus, estimates for underlying trends in crop yields are often derived from data at the regional or national level. However, the use of aggregated data for detrending may oversimplify analysis not only in cases of very high farm-level heterogeneity. For instance, CLAASSEN and JUST (2011) investigate the impact of aggregation on systematic and random yield variation; their finding from a large data set of non-irrigated corn in the US Corn Belt and Northern Plains suggest that both random as well systematic variation may be severely reduced by county-level aggregation. A more precise analysis of both study regions reveals that the largest share of systematic variation (“the trend component”) is within-country variation. Consequently, aggregated data poorly represent the farm yield variability and the risk faced on the farm unit level. Similarly, JUST and WENINGER (1999) conclude that “farm-specific estimation of the deterministic component is necessary to reflect heterogeneity of soil, production practices, and technology among farms”. However, this is only possible when long-term farm data are available, otherwise few observations may provoke severely biased results (ATWOOD et al., 2003).

Various approaches to yield trend estimations are present in the literature. For instance, BESSLER (1980), GOODWIN and KER (1998), and KER and GOODWIN (2000) use an autoregressive integrated moving average (ARIMA) model to account for carry-over effects from previous years, such as exceptionally dry or wet weather conditions. CHEN and MIRANDA (2006) use piecewise linear splines to model yield trends by allowing for two distinct linear trends, and by using nonlinear least squares to estimate model parameters and the breakpoint. Similarly, various piecewise linear regression models were tested by RONDANINI et al. (2012) and MOSS and SHONKWILER (1993) proposed a stochastic trend model. In most cases, however, polynomial regression is used (MIRANDA and GLAUBER, 1997; JUST and WENINGER, 1999) and trend parameters are estimated using OLS (SWINTON and KING, 1991). However, OLS is a non-robust estimation method and is thus highly sensitive to outliers, especially at the beginning and at the end of the yield series. Therefore, robust regression techniques that are not influenced by outlying observations, e.g. extremely low yield due to a drought event, have been suggested for crop yield data detrending (SWINTON and KING, 1991; FINGER, 2010b). SWINTON and KING (1991) compare OLS with six robust regression methods, and their findings suggest that robust regression may not lead to more reliable estimates. A later study by FINGER (2010b) revisits robust techniques by incorporating further developments in this field. FINGER (2010b) uses Monte Carlo simulations and shows that the robust MM estimator clearly outperforms OLS estimations for outlier-contaminated samples, and is similarly reliable for non-

contaminated samples. These two studies, however, focused on stochastic simulation of yield data rather than on the application of these methods to empirical data. To the authors' best knowledge, there is no study analysing (and validating) the capacity of this robust regression technique in the framework of empirical yield time series.

Besides using robust regression, another procedure to deal with high yield variability is proposed by BREUSTEDT et al. (2008). These authors include weather information as a proxy in their regression model to account for periods with both severe technological regress and adverse weather conditions. Thus, separating these two components by explicitly including a weather index as regressor, which represents the prevailing weather conditions during the crop growing season, may lead to more adequate trend estimations. As argued by BREUSTEDT et al. (2008), this procedure may be especially promising for longer yield time series and transition countries, where restructuring of economy also influences agricultural investments and outputs and accordingly, trends have to be described with e.g. quadratic or cubic polynomial functions. BREUSTEDT et al. (2008) applied OLS estimations in their analysis and did not explicitly evaluate the effect of the inclusion of additional information on trend estimates. Building upon this background, we contribute filling this gap in existing research by analysing the effect of this inclusion and evaluating whether a weather proxy reduces bias in trend estimations using robust MM estimator.

### 3 Data and Empirical Procedure

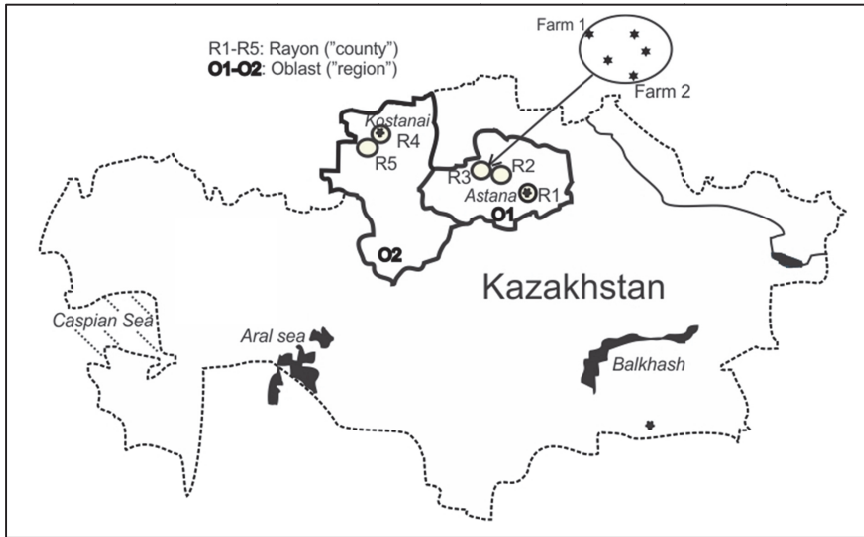
In this section, we first present the yield and weather data and then explain the model selection procedure applied in the study.

#### 3.1 Data

Agricultural production is an important sector for Kazakhstan (see e.g. TUBETOV et al., 2012; WORLD BANK, 2012) both, in terms of employment and countries economic performance. The main grain producing regions are situated in the northern parts and are predominantly extensive and low-cost production systems. Our data come from these regions and include 47 non-irrigated farms from five different counties ("rayons"). In addition, we use regional ("oblast") as well as national data (for an overview see Figure 1). Summary statistics are provided in Appendix A1. The structure of the data is as follows: rayons 1-3 are located in oblast 1 (Akmola oblast with capital Astana), and rayons 4 and 5 are situated in oblast 2 (Kostanai oblast). However, there are more rayons in each oblast than those used in the analysis. More precisely, there are 12 sample farms situated in rayon 1, 10 sample farms in rayon 2, 7 farms in rayon 3, 10 farms in rayon 4, and 7 farms in rayon 5. Once again, there are

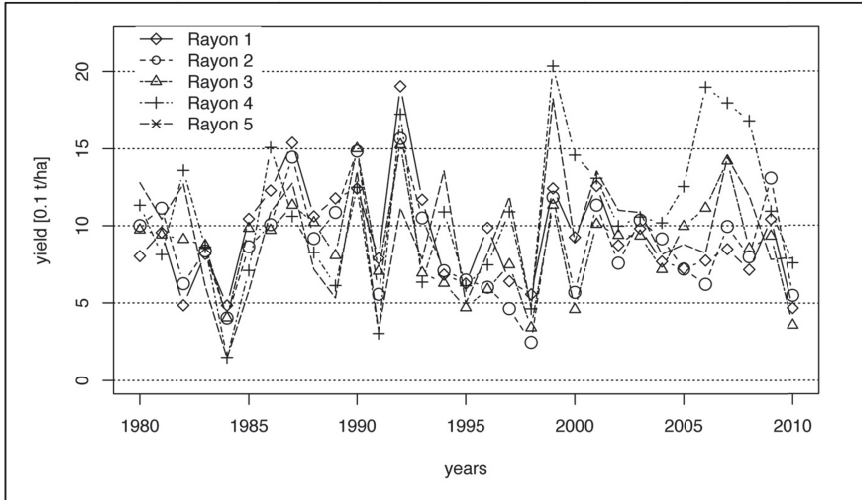
more farms than these sample farms in each rayon, and the average result of the sample farms may be different than the one on the rayon level. Farm and county' yield data were provided by regional statistical offices, and cover the period from 1980 to 2010. National and regional yields were obtained from the National Agency for Statistics of the Republic of Kazakhstan, as well as regional statistical offices. Our dataset represents 13 069 000 ha of wheat production, with a farm level average area under wheat of 16 000 ha.

**Figure 1. Map of Kazakhstan and study region**



Source: authors

Due to a high degree of soil heterogeneity, substantial variations of yield levels exist among the rayons, as well as between farms. Figure 2 shows the yield data of the five study rayons. The substantial year-to-year variation mainly due to droughts becomes apparent.

**Figure 2. Yield time series for the five study rayons**

Source: authors based on data from the regional statistical offices of Kazakhstan

The weather data were provided by the National Hydro-Meteorological Agency of Kazakhstan for five weather stations, each corresponding to a particular study rayon. We considered two different indices in our analysis: a hydrothermal coefficient by Selyaninov (Selyaninov index) and a cumulative rainfall index (CR). The Selyaninov index (SI) (MESHCHERSKAYA and BLAZHEVICH, 1996; DRONIN and KIRILENKO, 2008) is calculated as a sum of the ratios of cumulative rainfall,  $R$ , to the sum of daily average temperatures,  $T$ , for three sub-periods (3<sup>rd</sup> decade of May to July), which is the main growing season in Kazakhstan<sup>1</sup>:

$$SI = 10 \cdot \frac{R^{3rd\ dec.May}}{T^{3rd\ dec.May}} + \frac{R^{June}}{T^{June}} + \frac{R^{July}}{T^{July}}. \quad (1)$$

Higher values of the index represent favourable weather conditions, while lower index values refer to droughts. The cumulative rainfall index was calculated as the sum of precipitation [mm] from April to July:

$$CR = R^{April} + R^{May} + R^{June} + R^{July}. \quad (2)$$

<sup>1</sup> In general, the Selyaninov index is determined for daily average temperature above 10°C. We confined the index to the main growing season in Kazakhstan.



These two indices, SI for oblast 1 and CR for oblast 2, are used as a proxy for droughts, one of the main natural hazards for Kazakhstani grain production. An overview of the weather indices is provided in Appendix A2.

### 3.2 Empirical Procedure

To estimate the technological trend component, we apply Ordinary Least Squares (OLS) and the MM-estimator (Modified M-estimator), a robust regression technique proposed by YOHAI et al. (1991). The aim of robust regression methods is to provide unbiased estimations under situations, where non-robust methods, such as OLS, fail. This is for instance the case for outlier-contaminated data. The resistance of an estimator towards outliers can be described using the breakdown point. The MM-estimator has a breakdown point of 0.5, which means that MM can deal with an outlier contamination up to 50%, the highest possible value. In contrast, OLS has a breakdown point of 0, as one single outlier can have unconfined influence on the estimations. This is due to the different loss functions  $\rho$ , representing the residual weighting scheme: Whereas OLS minimizes a quadratic function, the MM estimator uses a less rapidly increasing function (bi-square re-descending score function).

The MM-estimator is based on an iterative algorithm which combines a highly robust but inefficient S-estimator with a highly efficient but non-robust M-estimator.<sup>2</sup> The M-estimator is defined as:

$$\hat{\beta}_M = \arg \min_{\beta} \sum_{i=1}^n \rho \left( \frac{r_i(\beta)}{\sigma} \right) \quad (3)$$

where  $r_i(\beta)$  denotes the regression residuals and the robust residual scale  $\sigma$  is used to standardize the residuals.

The S-estimator is given by:

$$\hat{\beta}_S = \arg \min_{\beta} \hat{\sigma}_s(r(\beta)) \quad (4)$$

<sup>2</sup> To improve the estimation of the starting values determined by the S-estimator and to allow for some improvements concerning the re-descending psi-function of the M-estimator, we use the specification *setting="KS2011"* in the function *lmrob* of the R-package *robustbase* (R DEVELOPMENT CORE TEAM, 2013). In cases of high parameters to sample size ratio, the S-scale estimate suffers bias, and KOLLER and STAHEL (2011) propose a "novel scale estimate based on the MM-estimate's residuals", called the "D-scale". While specifying *setting="KS2011"* in R, this D-estimation of scale step is included in the algorithm and referred to as the "SMDM" method (i.e. the combination of a S-estimation, M-estimation and D-estimation procedure).

The iterative algorithm starts with the S-estimator to provide initial residuals and a scale value ( $r_i$  and  $\sigma$ , equation 4). These initial estimates are then used to compute the first iteration of the regression coefficients ( $\hat{\beta}_M$ , equation (3)) using the M-estimator. These steps are reiterated until convergence is reached.

More information on the MM-estimator and the specifications are provided by KOLLER and STAHEL (2011), MARONNA et al. (2006) and YOHAI (1987).

Using these regression techniques, we construct two trend models. The first model, referred to as ‘reduced model’, does not account for weather effects (equation (5)), whereas the second model, the ‘full model’, includes weather effects as an additional regressor (equation (6)). The weather variable is used as a proxy<sup>3</sup> and we compare and analyse the estimated technological trend component of these two models. Both the full and the reduced models are tested for a trend up to a cubic polynomial degree by comparing the nested models. The most appropriate polynomial degree within each technique is determined using F-tests for OLS, and a robust Wald-type test for MM.<sup>4</sup> We considered cubic trends in our analysis because Kazakh wheat production was facing a phase of yield reductions due to disinvestment during the 1990s (during the transition period) that was enclosed by phases of technical improvements and significant yield increases until the late 1980s, and from the beginning of the 21<sup>st</sup> century. Against this background, we disregarded a higher polynomial degree, which, in addition, would severely increase the risk of overfitting due to the high parameter-to-sample ratio. To avoid model misspecification and not to undermine the ability of OLS for efficient estimations, we tested the statistical model assumption of independent and normally distributed error terms.<sup>5</sup>

---

<sup>3</sup> The p-values of the weather index are highly significant for all farms.

<sup>4</sup> We used a 5 % significance level for each test.

<sup>5</sup> To test for structural breaks we use three different methods: the Chow-test; the Rec-CUSUM test; and the Nyblom-Hansen test of the R package “strucchange”. The different tests do not agree on the farms having structural breaks, and we thus do not exclude farms from our sample. We found the following structural breaks for OLS: a) Full model: Chow-test: farm numbers 1 and 2; Rec-CUSUM test: farm number 14; Nyblom-Hansen test: farm number 25. b) Reduced model: Chow-test: farm number 24 and 26; no significant structural breaks for Rec-CUSUM and Nyblom-Hansen test. In addition, we test for autocorrelation using the Breusch-Godfrey test. The reduced and full model show significant autocorrelation for farm numbers 2 and 30. We examine the residuals with the autocorrelation function and partial autocorrelation function for lag  $k$ , and visualize the results using a correlogram. We specify an Autoregressive Moving Average model ARMA(1,0) for farm 2 and ARMA (0,1) for farm 30 using Generalized Least Squares. We re-estimate the model selection procedure and find the same trend as with OLS, where the residuals are assumed to be stochastically independent.

Reduced model

$$y_{it} = \beta_{i0}$$

$$y_{it} = \beta_{i0} + \beta_{i1} \cdot t$$

$$y_{it} = \beta_{i0} + \beta_{i1} \cdot t + \beta_{i2} \cdot t^2$$

$$y_{it} = \beta_{i0} + \beta_{i1} \cdot t + \beta_{i2} \cdot t^2 + \beta_{i3} \cdot t^3$$

F-test / Wald test (5)

Full model

$$y_{it} = \alpha_i \cdot weather_{ct} + \beta_{i0}$$

$$y_{it} = \alpha_i \cdot weather_{ct} + \beta_{i0} + \beta_{i1} \cdot t$$

$$y_{it} = \alpha_i \cdot weather_{ct} + \beta_{i0} + \beta_{i1} \cdot t + \beta_{i2}t^2$$

$$y_{it} = \alpha_i \cdot weather_{ct} + \beta_{i0} + \beta_{i1} \cdot t + \beta_{i2}t^2 + \beta_{i3}t^3,$$

F-test / Wald test (6)

where  $y_{it}$  indicates predicted yield of farm / county  $i$  at time  $t$ , and  $t$  is the time index with  $t = 1980, \dots, 2010$ ;  $\beta_{i0}$  represents the intercept for farm / county  $i$ ,  $\alpha_i$  is the parameter estimate for farm / county  $i$  and  $weather_{ct}$  stands for a weather index for each county  $c$  at time  $t$ .

Following GOODWIN and MAHUL (2004), we normalize the detrended yields to the last year  $t_{end}$ <sup>6</sup>. Thus, we add the error term of each year and farm to the detrended yield level at  $t_{end}$ ,  $\hat{y}_{t_{end}}$  as shown in equation (7) (reduced model). In addition, for the full model, we have to account for the weather index of the corresponding rayon, see equation (8).

$$y_t^{detrended} = \hat{y}_{t_{end}} + e_t \tag{7}$$

$$y_t^{detrended} = \hat{y}_{t_{end}} - \alpha \cdot weather_{t_{end}} + e_t + \alpha \cdot weather_t \tag{8}$$

Based on these detrended yields, we calculate their variance and refer to it as “remaining yield variance” to evaluate the effect of the aggregation level on a basic statistical measure of risk.

---

<sup>6</sup> More precisely we did not normalised our data to the year  $t_{end}$  but to the year  $t_{end-1}$ , since  $t_{end} = 2010$  was an extreme drought year and thus does not adequately represent today's yield level.

## 4 Results

We used different aggregation levels to analyse the effect of aggregation on trend estimations. More precisely, we compared the estimations based on single farms, rayons (counties), oblasts (regions) and national data (see Table 1). The number in the table represents the number of farms with the respective estimated trend. For instance, 8 farms of the 12 sample farms situated in rayon 1 (“12 farms - Rayon 1”) show no significant trend when using the reduced form model for OLS. The trends determined from single farm data show strong differences, varying from no trend up to a cubic trend. Rayon 5 shows no significant trend for all 7 sample farms located in this area for OLS and MM as well for the full and reduced model. On the higher aggregation level, i.e. the rayon and the oblast, no significant trend was found for either estimators. In contrast, the full model determines a significant quadratic trend for OLS and MM for rayon 4 and oblast 2, and a significant MM cubic trend for rayon 1. Across the different aggregation levels and for both models, MM and OLS determine a rather similar trend pattern.

**Table 1. Trend estimations for OLS and MM reduced form models, 4 aggregation levels**

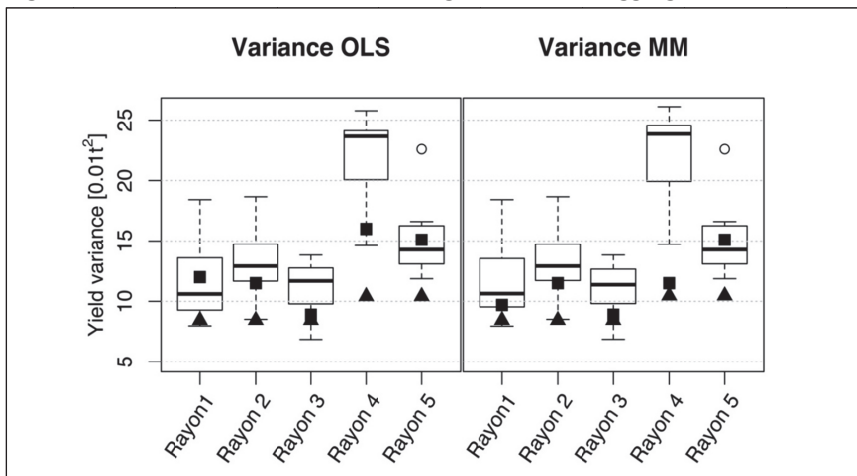
Aggregation level	OLS reduced / MM reduced				OLS full / MM full			
	No trend	Linear trend	Quadratic trend	Cubic trend	No trend	Linear trend	Quadratic trend	Cubic trend
National	-/-	-/-	1/1	-/-	*)			
Oblast 1	1/1	-/-	-/-	-/-	1/1	-/-	-/-	-/-
Oblast 2	1/1	-/-	-/-	-/-	-/-	-/-	1/1	-/-
Rayon 1 - Oblast 1	1/1	-/-	-/-	-/-	1/-	-/-	-/-	-/1
Rayon 2 - Oblast 1	1/1	-/-	-/-	-/-	1/1	-/-	-/-	-/-
Rayon 3 - Oblast 1	1/1	-/-	-/-	-/-	1/1	-/-	-/-	-/-
Rayon 4 - Oblast 2	1/1	-/-	-/-	-/-	-/-	-/-	1/1	-/-
Rayon 5 - Oblast 2	1/1	-/-	-/-	-/-	1/1	-/-	-/-	-/-
12 farms - Rayon 1	8/7	1/1	1/1	2/3	3/3	2/3	3/3	4/3
11 farms - Rayon 2	10/10	-/-	-/-	1/1	8/8	1/1	-/-	2/2
7 farms - Rayon 3	5/5	2/2	-/-	-/-	5/4	2/2	-/1	-/-
10 farms - Rayon 4	3/4	6/5	1/1	-/-	-/-	3/3	7/7	-/-
7 farms - Rayon 5	7/7	-/-	-/-	-/-	7/7	-/-	-/-	-/-

Note: numbers in the table indicate the number of farms with the respective trend estimation. \*) Not estimated because the national agricultural area is too large to be adequately represented by the weather information of a regional weather station.

Source: authors

Figure 3 shows the remaining yield variance, i.e. the variance after detrending for the farm-, rayon- and oblast-level for the full model. The boxplots represent the distribution of yield variances of the sample farms from corresponding rayons. Each farm has one yield variance estimate, and the sample size used to derive a boxplot varies with the number of sample farms in the corresponding rayon (i.e. between 7 and 12). The squares in Figure 3 refer to variances on the rayon level, and triangles refer to variances on the oblast level. The first three oblast variance values are identical, because rayons 1 to 3 are situated in the first oblast and the following two oblast variance values are identical, because rayons 4 and 5 are both situated in the second oblast. The variance is similar for the MM and OLS estimators, and for the reduced and full model (results for reduced model not shown). However, an increase in the aggregation level, i.e. from farm to rayon and to oblast, reduces yield variances. The median yield variance based on the farm level (given by the horizontal bar of the boxplot) is, except for rayon 5 (for OLS in addition rayon 1), higher than the one based on the rayon level and higher than one on the oblast level. However, the yield variance among farms are highly diverse, as seen by the spread of the boxplots. In rayon 4, the yield variance is outstanding, with a median variance of nearly 24 [0.01 t<sup>2</sup>]. These variance values even exceed the ones of rayon 5, where on the farm and the rayon level, no significant trend was determined and, as consequence for detrending, only a constant intercept term was removed.

**Figure 3. Yield variance after detrending: full model, 3 aggregation levels**



Note: the squares represent the variance on the rayon level, and the triangles denote the variance on the oblast level. T denotes tonnes.

Source: authors

Table 2 compares the results of the trend estimations determined by OLS for the reduced and full models. For about 30% of the farms (14 of the 47 farms), the trend estimations differed for the two models. Additionally, the trend estimations of the reduced form model have generally a lower polynomial degree than the ones of the full model. For instance, “no trend” was determined for 33 farms for the reduced model compared to 23 farms for the full model. Similarly, higher trend estimations, i.e. quadratic or cubic trends, were favoured for 16 farms for the full model in contrast to 5 farms for the reduced form model. Moreover, for “7 farms - Rayon 5”, no significant trend was determined for either model.

**Table 2. Trend estimation results for OLS full and reduced form model, farm-level data**

	OLS full / OLS reduced				OLS full and reduced comparison: Identical / no identical time trend
	No trend	Linear trend	Quadratic trend	Cubic trend	
12 farms - Rayon 1	3/8	2/1	3/1	4/2	7/5
11 farms - Rayon 2	8/10	1/0	0/0	2/1	9/2
7 farms - Rayon 3	5/5	2/2	0/0	0/0	7/0
10 farms - Rayon 4	0/3	3/6	7/1	0/0	3/7
7 farms - Rayon 5	7/7	0/0	0/0	0/0	7/0
Total	23/33	8/9	10/2	6/3	33/14

Note: numbers in the table indicate the number of farms with the respective trend estimation.

Source: authors

Table 3 summarizes the estimates of the robust MM estimator. Similar to OLS (Table 2), for about 30% of the farms the estimated time trends of the full and reduced model differed (i.e. 14 of the 47 farms). Likewise, lower polynomial trends are more frequently estimated with the reduced compared to the full model. For instance, the full model determined 16 farms with a quadratic or cubic trend, whereas the reduced form model identified only 6 farms having these trends.

**Table 3. Trend estimation results for MM full and reduced form model, farm-level data**

	MM full / MM reduced				MM full and reduced comparison: Identical / no identical time trend
	No trend	Linear trend	Quadratic trend	Cubic trend	
12 farms - Rayon 1	3/8	3/1	3/1	3/2	8/4
11 farms - Rayon 2	8/10	1/0	0/0	2/1	9/2
7 farms - Rayon 3	4/5	2/2	1/0	0/0	6/1
10 farms - Rayon 4	0/4	3/5	7/1	0/0	3/7
7 farms - Rayon 5	7/7	0/0	0/0	0/0	7/0
Total	22/34	9/8	11/2	5/3	33/14

Note: numbers in the table indicate the numbers of farms with the respective trend estimation.

Source: authors

In Table 4 the trend estimations of the full OLS model are compared to the reduced MM detrending model. For more than 30% of the farms (15 of 47 farms), the trend estimations differed. We again find a tendency for higher trend estimations for the OLS full model compared to the MM reduced model. For example the OLS full model determines 16 farms to have a quadratic or cubic polynomial degree compared to only 6 farms for the MM reduced model.

**Table 4. Trend estimation results for OLS full model and MM reduced form model, farm-level data**

	OLS full / MM reduced				OLS full and MM reduced comparison: Identical / no identical time trend
	No trend	Linear trend	Quadratic trend	Cubic trend	
12 farms - Rayon 1	3/7	2/1	3/2	4/2	7/5
11 farms - Rayon 2	8/11	1/0	0/0	2/0	9/2
7 farms - Rayon 3	5/4	2/2	0/1	0/0	6/1
10 farms - Rayon 4	0/6	3/3	7/1	0/0	3/7
7 farms - Rayon 5	7/7	0/0	0/0	0/0	7/0
Total	23/35	8/6	10/4	6/2	32/15

Note: numbers in the table indicate the numbers of farms with the respective trend estimation.

Source: authors

## 5 Discussion

In our investigation about an adequate representation of the technological trend component, we used a unique data set from farms, rayons (counties) and oblasts (regions) from northern Kazakhstan. For historic reasons, farm size in this region averages about 16 000 ha. This means that even for the smallest examined unit of this analysis, i.e. the single farm, yields are aggregated across these large farming areas. The trend estimations based on the farm-level data vary considerably among farms and regions. Hence, some farms had no significant trend, whereas others revealed a linear, quadratic or cubic trend pattern. However, for the higher aggregation level, i.e. the rayon and oblast level, especially for the reduced form model, no significant trend was determined. This indicates that by using farm-level data we either highly overestimated the technological trend component, or underestimated the trend effect by applying rayon or oblast data. The specific impact of aggregation is, besides the extent of aggregated hectares, dependent on the uniformity of weather conditions. In our case study region, the high yield volatility is a consequence of widely-spread droughts, thus reflecting a systemic risk component. This means that there is, with respect to the weather conditions, a rather high homogeneity among the farms and among the regions. Hence, aggregation will only slightly reduce the complexity of the yield pattern, as the main yield volatility does not originate from an idiosyncratic component. Another important presupposition for using aggregated data is a homogeneous pattern of technological change. The Kazakh agricultural system was subject to restructuring and privatisation, which especially during the transition period in the 1990s, lead to strong differences across farms with respect to factor endowment, resource availability and investment possibilities. Thus, the underlying patterns of the technological trend are heterogeneous, causing diverse patterns of trend estimate observations at the farm-level. Unifying the trends may thus lead researchers to model severe misspecifications, causing over- or underestimation of production risk. We evaluated this effect by comparing the remaining yield variance, i.e. the variance after detrending. We, therefore, compared the variance for trend estimations based on farm, rayon and oblast data, and found a decreasing variance with an increasing aggregation level of the trend estimation. This finding indicates that using aggregated data may severely underestimate the risk borne by Kazakh farmers.

Furthermore, we analysed and compared the use of the MM estimator, a robust detrending technique. Robust estimators give less weight to outliers in the data set. In contrast, non-robust estimators such as Ordinary Least Squares (OLS) are much more sensitive to outlying observations. This is especially pronounced for short yield series, which are predominantly used in risk assessment due to data constraints. In settings where yields vary substantially and trend estimations are not stable, extending or shortening the time series for a few years may shift the deterministic trend. In these



cases, using robust regression techniques in addition to OLS can reveal potential problems in data analysis. Observations that deviate from the relationship described by the majority of the data can be detected, and the influence of these outliers can be bounded. However, the Kazakh data contain many outlying observations because droughts are such a frequent event. Indeed, about one-third of the years can be referred to as drought or severe drought years (see also Figure 2). Thus, these extreme observations are part of the general pattern and can no longer be denoted as “outliers”. In addition, we had no extreme values at the very beginning or end of the time series, which would have amplified outlier effects. This may explain why we found only slight differences in the detrended yield estimations of MM and OLS. For our specific case study, the robust estimator did not seem to produce more reliable trend estimations. However, this finding cannot be generalised, and depends on the data set. For instance, none of the 7 farms located in rayon 5 exhibit any significant trend, and visualization shows a white noise sequence with a mean shift. Consequently, detrending results are stable and neither additional regressors nor the choice of the estimation technique have any influence on the results. This is contrary to farms from e.g. the first rayon, where trend patterns are more diverse, and the technique as well as the additional regressor have an effect on the trend estimations. Thus, the detrending technique has to be tailored to the particular data set.

Finally, we investigated the inclusion of a weather index as a regressor in detrending. We found that the trend estimations differed (i.e. different trend models have been found) for about one-third of the farms for the reduced and full model. Thus, our findings suggest that trend estimations require a simultaneous consideration of both, the effects of technology and the effect of weather. This is not only the case for the non-robust OLS estimator, but also for the robust MM estimator. It could be that the simultaneous occurrence of consecutive years of very low yields due to adverse weather events and technological decline due to disinvestments might constrain adequate trend estimation. The omitted but highly relevant weather variable provokes severe model misspecifications and leads to an omitted variable bias. In addition, depending on the length of the time series, weather variables might themselves exhibit a significant trend due to climate change. Ignoring such a trend by reducing the regression model to time variables might lead to inconsistent estimates of the trend parameters because the model residuals would not be independent from trend model regressors, i.e. time variable(s). In this case, the “weather” trend component will be captured in the technological trend component, thereby increasing the omitted variable bias. Thus, adding a weather variable representing prevailing weather conditions during the crop growing season as an additional regressor may help to reduce over- and underestimation of technological trends. Underestimating the deterministic trend may be the prevailing challenge in the case of Kazakh wheat production, since the non-inclusion of weather information led to lower trend estimations in our study.

## 6 Conclusions

Due to data constraints, many empirical applications apply aggregated yields such as county or regional data for risk analysis and actuarial applications. However, we show that for these highly heterogenic yield data, aggregation may oversimplify trend analysis and provoke biased results. Consequently, aggregated data poorly represent farm yield variability and the risk faced on the farm unit level. As shown in the literature, the choice of the estimation technique is also relevant for obtaining consistent estimates of technological trends, since deviations from OLS model assumptions such as outliers may have a substantial effect on trend estimation results. In addition to the non-robust OLS estimator, we applied the MM-estimator, a robust regression technique. This fills a gap in the literature by testing potential gains from using this estimator with observed instead of simulated crop yield data. Our finding suggests that the MM estimator did not produce more reliable trend estimations compared to OLS. This is due to the specific pattern of our data and cannot be generalised. About one-third of the yields are extreme and thus outlying observations are part of the general pattern. Consequently even the robust MM estimator requires additional information e.g. in our study in form of a weather index. By accounting for weather-related variance, the variance that is not related to the technological trend is reduced and consequently a more stable trend is estimated.

To summarise, the presented analysis contributes to an improved crop yield analysis for many developing and transition countries that face similar conditions. A careful trend analysis under such conditions is essential, because neglecting or misspecifications of trends lead to biased estimations and impede the development of powerful risk management instruments. Hence, a better assessment of farmers' yield risks allows the implementation of more efficient risk management strategies at the farm- and county level.

## References

- ATWOOD, J., S. SHAIK and M. WATTS (2003): Are crop yields normally distributed? A re-examination. In: *American Journal of Agricultural Economics* 85 (4): 888-901.
- BESSLER, D.A. (1980): Aggregated personalistic beliefs on yields of selected crops estimated using ARIMA processes. In: *American Journal of Agricultural Economics* 62 (4): 666-674.
- BOKUSHEVA, R., I. BEZLEPKINA and A.O. LANSINK (2009): Exploring farm investment behaviour in transition: The case of Russian agriculture. In: *Journal of Agricultural Economics* 60 (2): 436-464.
- BOKUSHEVA, R. and G. BREUSTEDT (2012): The effectiveness of weather-based index insurance and area-yield crop insurance: how reliable are ex post predictions for yield risk reduction? In: *Quarterly Journal of International Agriculture* 52 (2): 135-156.

- BRUSTEDT, G., R. BOKUSHEVA and O. HEIDELBACH (2008): Evaluating the potential of index insurance schemes to reduce crop yield risk in an arid region. In: *Journal of Agricultural Economics* 59 (2): 312-328.
- CHEN, S. and M.J. MIRANDA (2006): Modeling yield distribution in high risk countries: Application to Texas upland cotton. Contributed paper for presentation to American Agricultural Economics Association, Long Beach, California, 23-26 July, 2006.
- CLAASSEN, R. and R.E. JUST (2011): Heterogeneity and distributional form of farm-level yields. In: *American Journal of Agricultural Economics* 93 (1): 144-160.
- COOPER, J., M. LANGEMEIER, G. SCHNITKEY and C. ZULAUF (2009): Constructing farm level yield densities from aggregated data: Analysis and comparison of approaches. Contributed paper for presentation to Agricultural and Applied Economics Association, Milwaukee, Wisconsin.
- DOUGHERTY, C. (2011): *Introduction to Econometrics*. Oxford University Press, Oxford.
- DRONIN, N. and A. KIRILENKO (2008): Climate change and food stress in Russia: what if the market transforms as it did during the past century? In: *Climatic Change* 86 (1-2): 123-150.
- FINGER, R. (2010a): Evidence of slowing yield growth - The example of Swiss cereal yields. In: *Food Policy* 35 (2): 175-182.
- (2010b): Revisiting the evaluation of robust regression techniques for crop yield data detrending. In: *American Journal of Agricultural Economics* 92 (1): 205-211.
- (2012): Biases in farm-level yield risk analysis due to data aggregation. In: *German Journal of Agricultural Economics* 61 (1): 30-43.
- GOODWIN, B.K. and A.P. KER (1998): Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. In: *American Journal of Agricultural Economics* 80 (1): 139-153.
- GOODWIN, B.K. and O. MAHUL (2004): Risk modelling concepts relating to the design and rating of agricultural insurance contracts. World Bank Policy Research Working Paper WPS 3392. World Bank, Washington, DC.
- HAFNER, S. (2003): Trends in maize, rice, and wheat yields for 188 nations over the past 40 years: a prevalence of linear growth. In: *Agriculture, Ecosystems and Environment* 97 (1): 275-283.
- JUST, R.E. and Q. WENINGER (1999): Are crop yields normally distributed? In: *American Journal of Agricultural Economics* 81 (2): 287-304.
- KER, A.P. and B.K. GOODWIN (2000): Nonparametric estimation of crop insurance rates revisited. In: *American Journal of Agricultural Economics* 82 (2): 463-478.
- KOLLER, M. and W.A. STAHEL (2011): Sharpening Wald-type inference in robust regression for small samples. In: *Computational Statistics & Data Analysis* 55 (8): 2504-2515.
- LEBLOIS, A. and P. QUIRION (2013): Agricultural insurances based on meteorological indices: realizations, methods and research challenges. In: *Meteorological Applications* 20 (1): 1-9.
- MARONNA, R.A., R.D. MARTIN and V.J. YOHAI (2006): *Robust statistics, theory and methods*. Wiley Series in Probability and Statistics, Wiley. John Wiley & Sons Ltd., West Sussex, England.
- MARRA, M.C. and B.W. SCHURLE (1994): Kansas wheat yield risk measures and aggregation: A meta-analysis approach. In: *Journal of Agricultural and Resource Economics* 19 (1).

- MESHCHERSKAYA, A.V. and V.G. BLAZHEVICH (1996): The drought and excessive moisture indices in a historical perspective in the principal grain-producing regions of the former Soviet Union. In: *Journal Climate* 10 (10): 2670-2682.
- MIRANDA, M.J. and J.W. GLAUBER (1997): Systemic risk, reinsurance, and the failure of crop insurance markets. In: *American Journal of Agricultural Economics* 79 (1): 206-215.
- MOSS, C.B. and J.S. SHONKWILER (1993): Estimating yield distributions with a stochastic trend and nonnormal errors. In: *American Journal of Agricultural Economics* 75 (4): 1056-1062.
- OZAKI, V.A. and R.S. SILVA (2009): Bayesian ratemaking procedure of crop insurance contracts with skewed distribution. In: *Journal of Applied Statistics* 36 (4): 443-452.
- R DEVELOPMENT CORE TEAM (2013): *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.
- RONDANINI, D.P., N.V. GOMEZ, M.B. AGOSTI and D.J. MIRALLES (2012): Global trends of rapeseed grain yield stability and rapeseed-to-wheat yield ratio in the last four decades. In: *European Journal of Agronomy* 37 (1): 56-65.
- SKEES, J.R., J.R. BLACK and B.J. BARNETT (1997): Designing and rating an area yield crop insurance contract. In: *American Journal of Agricultural Economics* 79 (2): 430-438.
- SWINTON, S.M. and R.P. KING (1991): Evaluating robust regression techniques for detrending crop yield data with nonnormal errors. In: *American Journal of Agricultural Economics* 73 (2): 446-451.
- TANNURA, M.A., S.H. IRWIN and D.L. GOOD (2008): *Weather, technology, and corn and soybean yields in the U.S. corn belt*. Marketing and Outlook Research Report 01. Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.
- TUBETOV, D., O. MUSSHOF and U. KELLNER (2012): Investments in Kazakhstani dairy farming: a comparison of classical investment theory and the real options approach. In: *Quarterly Journal of International Agriculture* 51 (3): 257-284.
- WORLD BANK (2012): *Kazakhstan: Agricultural Insurance, Feasibility Study*. The World Bank, Washington, DC.
- YOHAI, V.J. (1987): High Breakdown-Point and High Efficiency Robust Estimates for Regression. In: *The Annals of Statistics* 15 (2): 642-656.
- YOHAI, V.J., W.A. STAHEL and R.H. ZAMAR (1991): *A procedure for robust estimation and inference in linear regression*. Springer-Verlag, New York.

## Acknowledgements

The authors appreciate the financial support of the Swiss National Science Foundation in the framework of the SCOPE-Program.

---

Contact author:

**Sarah Conradt**

ETH Zurich, Institute for Environmental Decisions (IED), Agricultural Economics

SOL C7, Sonneggstrasse 33, 8092 Zurich, Switzerland

e-mail: conradts@ethz.ch

## Appendix

### Appendix 1. Description of yield data

**Table A1. Summary statistics: yield data 1980-2010**

Aggregation level	Mean 0.1 t/ha <sup>a)</sup>	Min. 0.1 t/ha	Max. 0.1 t/ha	sd <sup>b)</sup>
12 farms - Rayon 1	8.9	0.2	24.0	3.8
11 farms - Rayon 2	8.8	0.8	21.0	3.8
7 farms - Rayon 3	8.3	1.2	19.3	3.4
10 farms - Rayon 4	10.7	0.9	25.6	5.1
7 farms - Rayon 5	9.2	0.3	22.1	3.9
Rayon 1	9.4	4.2	19.7	3.5
Rayon 2	9.4	2.8	16.2	3.4
Rayon 3	8.7	2.9	15.9	3.0
Rayon 4	9.8	1.5	17.6	4.1
Rayon 5	9.0	1.0	16.0	4.0
National	9.8	5.2	14.8	2.7

Note: <sup>a)</sup>Unit: tons per hectare. <sup>b)</sup>sd refers to standard deviation.

Source: regional statistical offices of Kazakhstan

### Appendix 2. Description of weather indices

**Table A2. Summary statistics: weather indices, 1980-2010**

Rayon	CR April-July, mm				SI 3 <sup>rd</sup> decade May-July			
	Mean	Min.	Max	CV <sup>a)</sup>	Mean	Min.	Max.	CV <sup>a)</sup>
1	141	94	215	0.26	0.72	0.26	1.38	0.38
2	133	33	234	0.34	0.71	0.14	1.57	0.47
3	127	46	267	0.38	0.65	0.22	1.74	0.55
4	164	83	269	0.35	0.87	0.30	1.93	0.50
5	148	70	297	0.39	0.75	0.22	1.82	0.48

Note: <sup>a)</sup>CV refers to coefficient of variation, CR denotes cumulative rainfall and SI stands for Selyaninov index.

Source: National Hydro-Meteorological Agency of Kazakhstan