Fixed Effects and Beyond

Bias reduction, Groups, Shrinkage, and Factors in Panel Data*

Stéphane Bonhomme
University of Chicago

Angela Denis
Banco de España

July 22, 2024

Abstract

Many traditional panel data methods are designed to estimate homogeneous coefficients. While a recent literature acknowledges the presence of coefficient heterogeneity, its main focus so far has been on average effects. In this paper we review various approaches that allow researchers to estimate heterogeneous coefficients, hence shedding light on how effects vary across units and over time. We start with traditional heterogeneous-coefficients fixed-effects methods, and point out some of their limitations. We then describe bias-correction methods, as well as two approaches that impose additional assumptions on the heterogeneity: grouping methods, and random-effects methods. We also review factor and grouped-factor methods that allow coefficients to vary over time. We illustrate these methods using panel data on temperature and corn yields, and find substantial heterogeneity across counties and over time in temperature impacts.

JEL codes: C10. C50.

Keywords: Panel data, fixed effects, coefficient heterogeneity.

*The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Banco de España or the Eurosystem.

1 Introduction

Fixed-effects methods are widely popular in applied practice. Examples include fixed effects for individuals, firms, cities, counties, countries, products, markets, years, days of the year, and times of the day, to cite a few. A notable example of the use of fixed effects is two-way fixed-effects estimation for difference-in-differences, which has become a leading method in applied economics.

Traditionally, researchers have used fixed effects to control for the presence of unit-specific heterogeneity. First-differencing, within-group, and quasi-differencing methods all attempt to difference-out the fixed effects. Increasingly, however, applied researchers have been viewing effects heterogeneity as a central focus of their work. In the presence of heterogeneity, a concern is that estimates based on constant-coefficients models, such as two-way fixed-effects estimators, recover difficult-to-interpret weighted averages of individual effects. A growing literature aims at determining whether those weights are positive and, if not, modifies the methods to estimate positively-weighted average effects, see De Chaisemartin and d'Haultfoeuille (2023) for a survey.

However, the focus on average effects makes only partial use of the data, and may mask important effects heterogeneity. Panel data, of the type used in many two-way fixed-effects applications, provides researchers with the opportunity to estimate the heterogeneity in effects. The goal of this paper is to review a variety of methods, developed in the past few decades, which can be used to estimate heterogeneous effects under suitable assumptions, thus going beyond averages or weighted averages.

To fix ideas, consider the following linear model for an outcome Y_{it} and a covariate X_{it}:

$$Y_{it} = \beta_{it} X_{it} + \alpha_{it} + \varepsilon_{it}, \quad i = 1, ..., n, \quad t = 1, ..., T. \tag{1}$$

In (1), α_{it} represents the level of the outcome, and β_{it} represents an heterogeneous "treatment effect"; i.e., the marginal effect of an increase in X_{it}. For example, when $X_{it} \in \{0, 1\}$ is binary, β_{it} is equal to the difference in outcomes between two hypothetical values $X_{it} = 1$ and $X_{it} = 0$. For simplicity here we abstract from the presence of other covariates, although those are often present in applications.

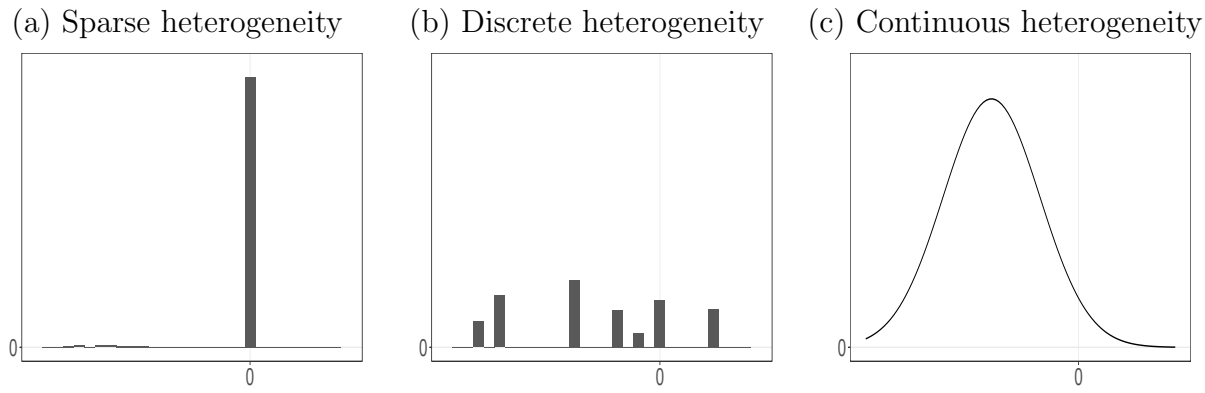
Two-way fixed-effects estimation is based on three key assumptions in model (1). The first assumption is that

$$\alpha_{it} = \alpha_i + \delta_t, \tag{2}$$

which is commonly referred to as a "parallel trends" assumption. The second assumption is that the coefficient of X_{it} is constant,

$$\beta_{it} = \beta. \tag{3}$$

Figure 1: Examples of time-invariant heterogeneity



The third assumption is that errors ε_{it} are mean independent of $X_{i1}, ..., X_{iT}$ and have mean zero. While this strict exogeneity assumption is empirically restrictive, it is commonly imposed in applications and we will maintain it in this paper.

A popular line of research in recent years has been to maintain the parallel trends specification for α_{it} in (2), while fully relaxing (3) by allowing β_{it} to be heterogeneous in unrestricted ways. Influential contributions include Goodman-Bacon (2021), Callaway and Sant'Anna (2021), De Chaisemartin and d'Haultfoeuille (2020), and Sun and Abraham (2021). There is also recent work relaxing the parallel trends assumption (2) (see Rambachan and Roth, 2023). The chief goal of these approaches is to learn about certain weighted averages of the β_{it}'s.

If the effects β_{it} are fully unrestricted, there is no way to estimate them consistently. In that case, the best researchers can hope for is to estimate some average or weighted average of effects. However, researchers may be willing to impose certain assumptions on β_{it}, while not restricting them to be constant as in (3). Doing so makes it possible to learn about how β_{it} varies across individual units and over time.

Consider first the case where heterogeneous effects $\beta_{it} = \beta_i$ do not vary over time. Then, the parameters β_i are *fixed effects* that can be estimated given a long enough panel. When the time dimension is short or moderate, however, estimates of β_i tend to be noisy. A pressing question in practice is whether the dispersion in estimates of β_i reflects actual heterogeneity, or whether it is due to sample noise; i.e., to the fact that each β_i is estimated with error.

We will review various approaches that address the issue of noise. The first class of methods do not impose further restrictions on β_i, but acknowledge the presence of noise in the estimates. Exact and approximate *bias-correction* methods have been developed in recent years and are

now well understood and applicable (e.g., Hahn and Newey, 2004, Arellano and Hahn, 2007).

However, bias-correction methods suffer from some of the same issues as fixed effects. In practice, due to low variability in covariates, the fixed effects cannot be calculated for some units (which get automatically dropped from the sample). In addition, fixed-effects methods and their bias-corrected counterparts are based on unit-by-unit estimation, which does not benefit from any pooling in the cross-section.

These issues can be alleviated if one is willing to impose some assumptions on β_i. In cross-sectional settings, a common assumption is to assume that the β_i's are sparse, so most of the β_i values are equal to zero or to a common value (see Figure 1 panel (a) for an illustration). However, such an assumption might not be a plausible model to describe parameter heterogeneity in panel data applications.

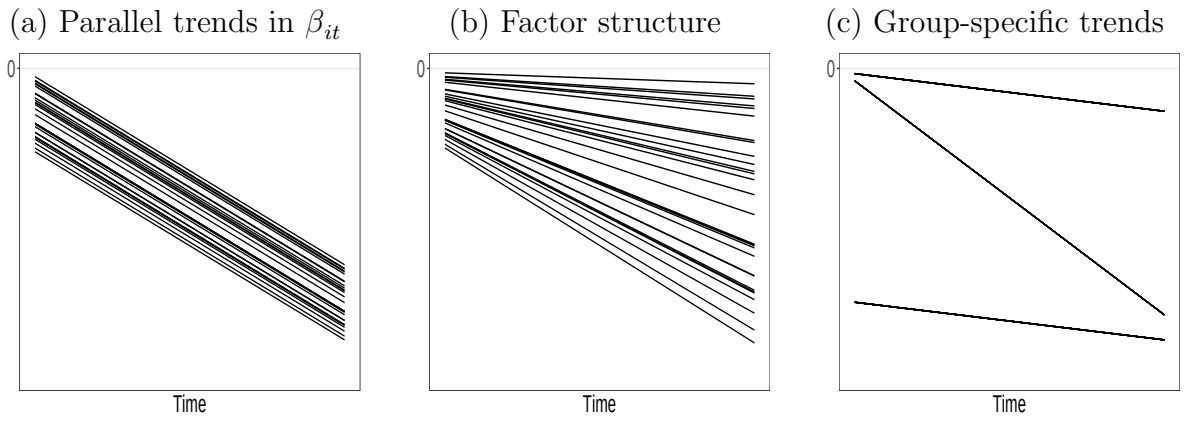
We will review methods based on two alternative assumptions: that β_i are grouped (see Figure 1 panel (b)), or that they follow some parametric or semi-parametric distribution (see Figure 1 panel (c)). Both *grouping methods* based on discrete heterogeneity (e.g., Bonhomme and Manresa, 2015), and *random-effects methods* with their connections to empirical-Bayes shrinkage (e.g., Efron, 2012, Gu and Koenker, 2017), can offer meaningful reduction in noise and improve the quality of heterogeneity estimates.

Consider next the case where β_{it} varies over time. In many applications, allowing for variation over time is empirically important. A common assumption is that the $n \times T$ matrix with elements β_{it} has low rank, which corresponds to a *factor structure*. A simple example is represented in Figure 2 panel (a), which corresponds to an additive model of β_{it} as the sum of a unit fixed-effect and a time fixed-effect. However, factor structures can accommodate more general patterns of heterogeneity. A simple example is shown on Figure 2 panel (b), which features a one-factor specification. Moreover, the intercept α_{it} can also be modeled using a factor structure, thus relaxing the parallel trends assumption (2). We will review methods for interactive fixed-effects regressions and its generalizations, building on Bai (2009).

A drawback of factor methods is that they do not address the noise issue in fixed-effects estimation. Indeed, compared to the case where β_i is time-invariant but otherwise unrestricted, factor models depend on numerous additional parameters. This can make their use problematic in settings where the time dimension is not sufficiently large. An alternative approach relies on the assumption that β_{it} follows some group-specific time trend, as represented in Figure 2 panel (c). We will review such *grouped-factor methods* and highlight their ability to capture time variation in unit heterogeneity.

We will illustrate a number of these methods by estimating how temperature affects agri-

Figure 2: Examples of time-varying heterogeneity



cultural output, specifically corn production, in the US. A large literature, including Deschênes and Greenstone (2007), Schlenker and Roberts (2009), and Burke and Emerick (2016) estimates temperature impacts based on some variants of two-way fixed-effects. Estimates of temperature impacts are key inputs to calculations of the costs of climate change, and documenting how these impacts vary across counties in the US and over time is important to inform such calculations, as recently demonstrated by Keane and Neal (2020). We will report estimates based on various methods, including bias-corrected estimates, grouped and grouped-factor methods, and random-effects methods, allowing for effects heterogeneity across space but also over time.

While this paper focuses on bias correction, grouping, random-effects, and factor approaches, it omits other related methods. Many of the other methods we do not review focus on average effects. For example, there is a large earlier literature on slope heterogeneity in panel data (e.g., Hsiao and Pesaran, 2008, Wooldridge, 2005). The recent literature on treatment effects estimation using panel data pursues a related goal, see Arkhangelsky and Imbens (2023) and the recent survey in Arkhangelsky and Imbens (2024). We also do not cover work using synthetic control and synthetic difference-in-differences methods (e.g., Abadie, Diamond, and Hainmueller, 2010, Arkhangelsky, Athey, Hirshberg, Imbens, and Wager, 2021).

The outline of the rest of the paper is as follows. In Section 2 we present our empirical illustration to quantify the impact of temperature on corn yields. In Section 3 we describe fixed-effects methods, and review approaches to bias correction in Section 4. In Sections 5 and 6 we review grouped fixed-effects and random-effects methods. Lastly, in Section 7 we review factor and grouped-factor methods to allow for time-varying heterogeneity, and we conclude in Section 8.

2 An illustrative application: agriculture and the weather

To illustrate the methods, we will use US panel data to study the relationship between temperature and agricultural output.

2.1 Model and objectives

Consider panel data on counties $i = 1, ..., n$ and years $t = 1, ..., T$. Let Y_{it} denote corn yields, which are our measure of agricultural output. Let X_{it} denote a measure of temperature. We will estimate various versions of the following model:

$$Y_{it} = \beta_{it} X_{it} + \alpha_{it} + W_{it}'\gamma + \varepsilon_{it}, \tag{4}$$

where W_{it} includes a set of control variables (i.e., precipitation and state-year indicators) and ε_{it} is a mean-zero error term.

A special case of (4) is

$$Y_{it} = \beta X_{it} + \alpha_i + \delta_t + W_{it}'\gamma + \varepsilon_{it}. \tag{5}$$

A common approach in the literature relies on within-county regressions based on (5) to estimate short-run temperature impacts. Estimates of β inform the projected costs of climate change scenarios (e.g., Deschênes and Greenstone, 2007, Dell, Jones, and Olken, 2014). Note that (5) is a two-way fixed-effects regression model.

The methods we review in this paper will allow us to generalize (5) and estimate various versions of model (4). Our main goal when relaxing (5) is to allow β_{it} to be heterogeneous across counties and over time. This heterogeneity is of substantive interest for climate change calculations (Keane and Neal, 2020). We will document heterogeneity in temperature impacts across space and over time.

2.2 Data and preliminary evidence

We use two sources of data. For the weather variables, we use a balanced dataset of US counties since 1950, constructed from daily weather records.[1] We aggregate the data at the county-year level. For the output variables we focus on corn yields, and use an unbalanced panel of US

[1] The records are based on PRISM data and available on Wolfram Schlenker's website.

counties for the period 1950–2005 constructed by Burke and Emerick (2016), which comes from the US Department of Agriculture's National Agricultural Statistics Service.[2]

Following the literature, we focus our analysis of corn yields, measured in bushels per acres planted, on counties east of the 100th meridian where agriculture is primarily rainfed as opposed to irrigated, which account for the vast majority of US corn production. We randomly remove one-third of the counties per state for the years 2001–2005, and will use these observations as a hold-out sample to perform prediction exercises (see Section 7). The sample for analysis excludes the hold-out sample, drops observations without yields or corn area, and only keeps counties that have at least 10 years of valid data. This sample has 2,253 counties and 104,149 observations.

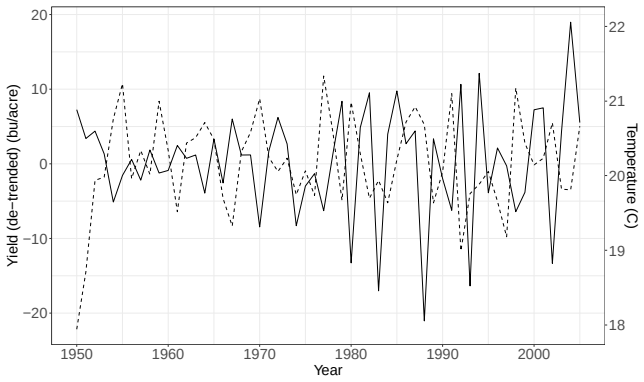
As our main weather variable we focus on temperature, measured as average daily degree Celsius above zero from April 1st to September 30th, which is the growing season for corn. It is worth noting at the outset that the within–county variance in temperature is only 5.7% of the total variance. Although temperature is the main focus of our analysis, we also control for average daily precipitation during the growing season each year, in millimeters per day. In Appendix Table A1 we report descriptive statistics on the weather variables, corn yields, and the area of corn production.

Our focus on the marginal impact of one additional degree on yields misses several important margins that have been emphasized in the literature on the agricultural impacts of changing temperatures. Previous work has shown that the impact of temperature on yields is nonlinear (e.g., Schlenker and Roberts, 2009). Such nonlinearity is not captured by the homogeneous specification (5). However, specifications allowing for β_{it} in (4) to be heterogeneous and correlated with temperature do allow temperature effects to depend on temperature levels. A more general specification would allow for square or high-order terms in temperature X_{it} in (4). Another restrictive feature of our main equation is the absence of a role for extreme temperatures and other climatic events, which have been shown to be important (e.g., Miller, Chua, Coggins, and Mohtadi, 2021). A third feature is the absence of a role for farmers' adaptation to changing temperatures (Burke and Emerick, 2016, Keane and Neal, 2020).

We first plot temperature and de-trended yields over the sample period. Figure 3 shows clear evidence of a negative correlation between the two, suggesting that higher temperatures lead to lower output. To account for heterogeneity over time and across counties in the *level*

[2]We downloaded the data from the AEA's website. The data period in the sample used by Burke and Emerick (2016) ranges until 2010. However, the number of counties decreases by more than 40% for the last 5 years. For this reason, we do not include those years in our analysis.

Figure 3: Temperature and de-trended yields



Notes: un-weighted statistics across counties. Yields are measured in bushels per acre. Temperature is measured in average daily degree Celsius above zero during the growing season. De-trended yields are shown in solid (left axis), temperature in dashed (right axis).

Table 1: Regression estimates with constant coefficients

	(1)	(2)	(3)	(4)
Temperature	-3.924	-3.488	-3.875	-6.600
	(0.700)	(1.285)	(0.818)	(0.880)
Precipitation	5.664	7.244	3.097	1.991
	(1.578)	(1.054)	(0.519)	(0.380)
Observations	104,149	104,149	104,149	104,149
County FE	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
State-year FE	No	No	No	Yes

Notes: un-weighted regressions of corn yields on temperature and precipitation. Standard errors are clustered at the state level.

of yields and temperature (yet not for heterogeneity in temperature *effects*), it is common to estimate some version of equation (5).

We report estimates of such regressions in Table 1, for a variety of controls that always include precipitations, and, in the richest specification in column (4), also include county and state-year fixed-effects. In that specification, parallel trends are required to hold within state, but state-specific trends are unrestricted. All the specifications require the temperature effects β to be constant across counties and over time.

The estimates in Table 1 show that an increase in one degree per day is associated with a drop in yields of 6.6 bushells per acre, which represents a 9% drop relative to the average level of the yields. This estimate is nearly twice as large as estimates that do not account for state-year fixed-effects.[3]

3 Fixed-effects

In this section we first review traditional fixed-effects methods.

3.1 Model and assumptions

We start by assuming that effects heterogeneity β_{it} is constant over time, equal to β_i, and estimate various versions of the following two-way fixed-effects model with slope heterogeneity:

$$Y_{it} = \beta_i X_{it} + \alpha_i + \delta_t + W_{it}'\gamma + \varepsilon_{it}. \tag{6}$$

In the following we will focus on the case where X_{it} is scalar, since this corresponds to the case of our application. However, the methods are easily extended to the case where X_{it} is multivariate and β_i is a vector.[4]

Note that assuming $\beta_{it} = \beta_i$ restricts the $n \times T$ matrix

$$\boldsymbol{B} = \begin{pmatrix} \beta_{11} & \beta_{12} & ... & \beta_{1T} \\ \beta_{21} & \beta_{22} & ... & \beta_{2T} \\ ... & ... & ... & ... \\ \beta_{n1} & \beta_{n2} & ... & \beta_{nT} \end{pmatrix} \tag{7}$$

to have identical columns. However, the fixed-effects approach leaves effects heterogeneity unrestricted across units. For example, it allows for situations where effects heterogeneity β_i is correlated across counties, and is correlated with temperature and precipitation in all counties in arbitrary ways.

In model (6), ordinary least squares (OLS) regression gives the fixed-effects estimators

$$\widehat{\gamma}, \widehat{\delta}_1, ..., \widehat{\delta}_T, \widehat{\alpha}_1, ..., \widehat{\alpha}_n, \widehat{\beta}_1, ..., \widehat{\beta}_n.$$

[3] Consistently with the literature, we cluster standard errors at the state level. Since there are only 31 states, we also computed Driscoll-Kraay standard errors as a robustness check, see Appendix Table B2.

[4] A practically important case where multivariate β_i's arise is when X_{it} includes lags of a covariate of interest. However, allowing for multivariate β_i's is more demanding than only allowing for scalar heterogeneity. The issues with fixed-effects estimation that we will highlight below become even more salient in the multivariate case.

We will focus on the coefficients $\widehat{\beta}_i$, which are estimates of β_i, for $i = 1, ..., n$. We now list three common assumptions under which we will review some properties of these estimators.

Assumption 1. *The matrix of regressors in (6) has full rank.*

Assumption 1 requires sufficient variation in the regressors. In models with slope and intercept heterogeneity this requires X_{it} to vary over time. To see this, consider the following simplified version of model (6), without additional covariates and time effects:

$$Y_{it} = \beta_i X_{it} + \alpha_i + \varepsilon_{it}. \tag{8}$$

In model (8) we have

$$\widehat{\beta}_i = \frac{\sum_{t=1}^{T} \left(X_{it} - \overline{X}_i\right)\left(Y_{it} - \overline{Y}_i\right)}{\sum_{t=1}^{T} \left(X_{it} - \overline{X}_i\right)^2}, \tag{9}$$

where $\overline{Z}_i = \frac{1}{T}\sum_{t=1} Z_{it}$ denotes the unit-specific average of any random vector Z_{it}. The denominator in (9) is proportional to the within-county variance of X_{it}. Good behavior of $\widehat{\beta}_i$, in the sense of a low variance, will require sufficient variability of X_{it} over time.

Assumption 2. *The covariates $X_i = (X_{i1}, ..., X_{iT})'$ and $W_i = (W_{i1}', ..., W_{iT}')'$ are strictly exogenous, in the sense that*

$$\mathbb{E}[\varepsilon_{it} \mid X_i, W_i] = 0 \text{ for all } i = 1, ..., n \text{ and } t = 1, ..., T.$$

Strict exogeneity restricts ε_{it} to be mean independent not only of past and current covariates, but also of future covariates. This is a strong assumption that rules out the presence of feedback from past outcomes to future covariates. See Arellano (2003b) for discussions of this assumption. Nevertheless, in the context of our application, strict exogeneity of temperature and precipitation may be plausible. Under Assumption 2, the $\widehat{\beta}_i$'s are unbiased:[5]

$$\mathbb{E}\left[\widehat{\beta}_i\right] = \beta_i.$$

Assumption 3. *Observations are independent across i.*

[5]Although we will maintain Assumption 2 throughout the paper for simplicity, fixed-effects and grouped fixed-effects methods remain theoretically justified as n and T tend to infinity under sequential exogeneity, that is,

$$\mathbb{E}[\varepsilon_{it} \mid X_{i1}, ..., X_{it}, W_{i1}, ..., W_{it}] = 0 \text{ for all } i = 1, ..., n \text{ and } t = 1, ..., T,$$

even though the $\widehat{\beta}_i$'s are no longer unbiased in that case. Note that, in contrast with strict exogeneity, sequential exogeneity allows future values $X_{i,t+h}$ (for $h \geqslant 1$) to correlate with past errors ε_{it}. See for example Fernández-Val and Lee (2013).

We will impose Assumption 3 for simplicity. Note that the controls W_{it} may include time effects, and they will include state-time indicators in our application. The inclusion of time effects captures some sources of cross-sectional dependence. Moreover, inference methods allowing for common shocks across individual units, in addition to serial correlation within units, have been developed in recent years, see in particular Andrews (2005), Kuersteiner and Prucha (2013), and Kuersteiner and Prucha (2020).

3.2 Mean and variance of fixed effects

Given Assumptions 1, 2 and 3, we now review some properties of fixed-effects estimators. Let

$$\widehat{\mathbb{E}}\left[\widehat{\beta}\right] = \frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}_i \tag{10}$$

denote the mean of the fixed-effects estimates, sometimes referred to as the *mean-group* estimator. Suppose that Assumptions 1 and 2 hold, so in particular covariates are strictly exogenous. Then the mean-group estimator (e.g., Chamberlain, 1992, Pesaran and Smith, 1995) is an unbiased estimator of the average of the effects (or average treatment effect)

$$\widehat{\mathbb{E}}\left[\beta\right] = \frac{1}{n}\sum_{i=1}^{n}\beta_i,$$

since we have

$$\mathbb{E}\left(\widehat{\mathbb{E}}\left[\widehat{\beta}\right]\right) = \widehat{\mathbb{E}}\left[\beta\right].$$

Moreover, under suitable regularity conditions, the mean-group estimator is also consistent for the average of the effects. Consistency holds as the cross-sectional size n tends to infinity, irrespective of whether T is fixed or tends to infinity. That is, as n tends to infinity,

$$\widehat{\mathbb{E}}\left[\widehat{\beta}\right] = \widehat{\mathbb{E}}\left[\beta\right] + o_p(1).$$

The conditions for consistency of the mean-group estimator require sufficient variability of X_{it} over time. In practice, mean-group estimates are sensitive to the covariates of *some* individual units exhibiting low variation over time. Graham and Powell (2012) point out this issue in the context of heterogeneous-coefficients panel data models with continuous covariates, and propose a trimming strategy for consistent estimation. Notice that the requirement for sufficient time variation is already apparent in Assumption 1.[6]

[6]To illustrate, consider model (8) with a binary covariate and $T = 2$. In that case, identification of the average $\frac{1}{n}\sum_{i=1}^{n}\beta_i$ requires $X_{i1} \neq X_{i2}$ for *all* individual units $i = 1, ..., n$ (i.e., that all units be so-called "movers"). If now X_{it} is continuous, identification only requires $X_{i1} \neq X_{i2}$ to hold almost surely, yet the mean-group estimator may be ill-behaved when a mass of individuals have $X_{i1} \approx X_{i2}$ (i.e., when there are "near stayers").

Next, let

$$\widehat{\text{Var}}\left(\hat{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_i - \hat{\mathbb{E}}\left[\hat{\beta}\right]\right)^2 \tag{11}$$

denote the sample variance of the fixed-effects estimates. Under the same assumptions, the sample variance is *biased* for the sample variance of the effects β_i,

$$\widehat{\text{Var}}\left(\beta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_i - \hat{\mathbb{E}}\left[\beta\right]\right)^2.$$

Specifically, we have, under Assumptions 1 and 2,

$$\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\beta}\right)\right] = \widehat{\text{Var}}\left(\beta\right) + \text{Bias},$$

where the bias term is positive. Intuitively, the dispersion in the fixed-effects estimates reflects not only the dispersion in the true effects, but also some additional dispersion due to noise.

To derive the expression for the bias term, consider for simplicity model (8) without covariates. Using Assumptions 1, 2 and 3 we obtain

$$\text{Bias} = \left(1 - \frac{1}{n}\right)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T}\sum_{s=1}^{T}\left(X_{it} - \overline{X}_i\right)\left(X_{is} - \overline{X}_i\right)\varepsilon_{it}\varepsilon_{is}}{\left[\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2\right]^2}\right]. \tag{12}$$

We notice in (12) that the bias does *not* go away as n tends to infinity. However, it tends to decrease as T grows, and to vanish as T tends to infinity. A typical order of magnitude is $O(1/T)$, meaning that the bias is twice as small when the panel length double. We also see that the magnitude of the bias is affected by the lack of variability in X_{it} over time, as reflected by the sample variance of X_{it} in the denominator in (12). Lastly, we see the impact of the noise through the presence of ε_{it}. Everything else equal, a larger variance of ε_{it} tends to increase the bias, and a higher persistence among ε_{it}'s over time tends to lead to larger bias as well.

3.3 Four issues with fixed effects

In models with heterogeneous parameters such as (6), fixed-effects methods are useful tools to quantify the heterogeneity in individual responses. However, fixed-effects estimators suffer from four main issues that limit their appeal for applied practice.

The *first issue* is apparent from the analysis of the variance that we have just reviewed: fixed-effects are noisy, which tends to bias the parameters of interest, such as the dispersion in individual treatment effects. Should one trust the dispersion of fixed-effects estimates, or are those partly capturing spurious heterogeneity?

The *second issue* is that, in many data sets, some fixed effects cannot even be calculated, due to lack of variation in covariates for some individual units. Technically, this reflects a violation of Assumption 1, which is a common occurrence in applications. This issue gets sometimes unnoticed in empirical work, since statistical software often automatically drops observations and their associated coefficients when those are not identified.

The *third issue* is that the fixed-effects approach relies on i-by-i estimation, without aiming at exploiting any information across individual observations. As an illustration, consider model (8) without additional covariates. We see that each estimate $\widehat{\beta}_i$ in (9) only depends on the observations Y_{it} and X_{it} corresponding to the individual unit i, without any sort of pooling across individual observations.

The *last issue* with fixed-effects is an obvious one: fixed effects are fixed over time. It is not possible to adopt a fixed-effects approach to estimate unrestricted β_{it} parameters. However, in applications, it is often appealing to allow for time variation in individual responses.

In the subsequent sections we will review various approaches that aim at dealing with these four issues with fixed-effects estimation. Before doing so, we start by reporting fixed-effects estimates in our application.

3.4 Illustration

We estimate model (6) with county-specific coefficients β_i, where we control for state-year fixed effects and county fixed effects, in addition to precipitation. Column (2) in Table 2 shows that the average $\widehat{\beta}_i$ in the sample is -7.9.[7] This is larger than the estimate based on a homogeneous specification (compare with Column (1)). Moreover, the average masks considerable heterogeneity. In some counties, temperature has small effects on corn yields: the estimate is -2.4 at the 90th percentile. However, in other counties, the effects are much more detrimental, as evidenced by the value of the 10th percentile (-14.7). The standard deviation of $\widehat{\beta}_i$ is 5.0.

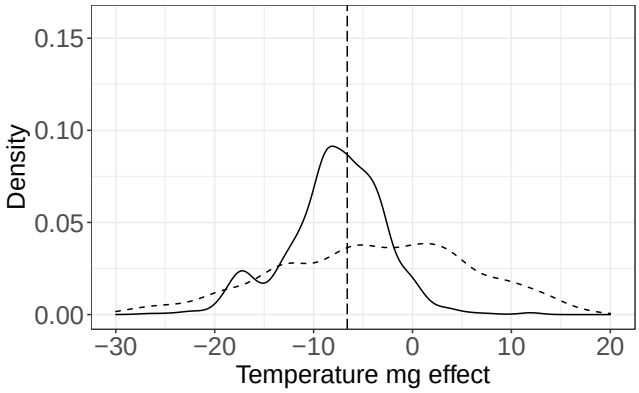
County heterogeneity in temperature impacts can be seen in the kernel estimate of the density of $\widehat{\beta}_i$ shown in solid line in the left panel of Figure 4. While effects are overwhelmingly negative, the estimates show a tick left tail of large negative impacts. To assess the extent of spatial heterogeneity, in the right panel of Figure 4 we plot the estimates $\widehat{\beta}_i$ on a map of the US.[8]

[7] In this average, counties are weighted by the total corn area in years they appear in the sample. We proceed similarly for other quantities such as variances and percentiles of effects.

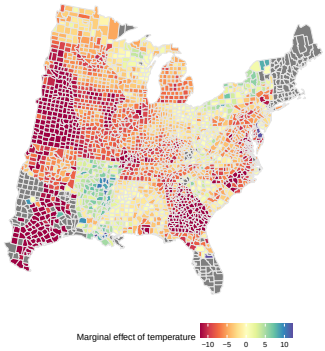
[8] While we do not use weights in the regressions, to present the results we weight all estimates by corn area

Figure 4: Fixed-effects estimates with heterogeneous coefficients

(a) Marginal density (short panel in dashed)



(b) Spatial distribution



Notes: panel (a) shows the weighted density of marginal effects across counties and years estimated in the long panel 1950–2005 (solid line), versus the one estimated in the short panel 1990–2005 (dotted line) for the same sample of counties. Panel (b) plots the marginal effects per county estimated using the full sample.

In the appendix we report two sets of robustness checks to assess how our findings are affected by some specification changes. Some authors use log yields instead of yield levels as the left-hand side variable. In Appendix B.3 we show estimates under this specification. Moreover, some authors rely on different measures of temperature, which capture temperature under normal conditions ("growing degree days") and extremely high temperatures ("killing degree days"), see for example Keane and Neal (2020). In Appendix B.4 we report estimates based on such a specification. We find that, relative to the marginal effects coefficients we focus on in this paper, extreme temperature impacts ("killing degree days") are differently distributed across space, see Appendix Figure B7. It is important to keep these differences in mind when interpreting our findings in light of the literature.

A key question we ask is: how should one interpret the heterogeneity in these estimated impacts? Does it reflect true heterogeneity in β_i or sample noise? As a first look at this question, we estimate the same model on a shorter panel, from 1990 to 2005. We then plot a kernel estimate of the density of the $\widehat{\beta}_i$'s estimated on this subsample, in dashed line in the left panel of Figure 4. We see a density of impacts that is much wider than in the full sample, with large proportions of negative but also positive temperature impacts. While it may be that the last period 1990-2005 experienced very different temperature impacts compared to the full sample, it is also possible that the difference between the two densities is mechanically due to

in the county. In Appendix B.2 we show un-weighted estimates for comparison.

the effect of sample noise. In the next section, we review bias-correction methods that aim at reducing the impact of the noise on fixed-effects estimates.

4 Bias-correction methods in fixed effects

4.1 Exact bias correction

A number of methods are available to either fully correct for bias in fixed-effects estimates, or to at least partially correct for it. We first discuss exact bias-correction methods.

Consider the variance of fixed-effects estimates in model (8) without additional covariates. The bias term in (12) depends on the variance-covariance matrix of errors ε_{it}. As an example, in the case where $\varepsilon_{it} \,|\, X_{i1}, ..., X_{iT} \sim iid(0, \sigma_\varepsilon^2)$ we have

$$\text{Bias} = \sigma_\varepsilon^2 \left(1 - \frac{1}{n}\right) \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}(X_{it} - \overline{X}_i)^2\right)^{-1}\right].$$

An unbiased and consistent estimator of the bias can then be constructed as

$$\widehat{\text{Bias}} = \widehat{\sigma}_\varepsilon^2 \left(1 - \frac{1}{n}\right) \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}(X_{it} - \overline{X}_i)^2\right)^{-1},$$

where

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{n(T-2)}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \widehat{\beta}_i X_{it} - \widehat{\alpha}_i\right)^2$$

is the degrees-of-freedom-corrected OLS estimate of the error variance.

A bias-corrected, exactly unbiased variance estimate is easily obtained as

$$\widehat{\text{Var}}^{\text{BC}}(\beta) = \widehat{\text{Var}}\left(\widehat{\beta}\right) - \widehat{\text{Bias}},$$

which satisfies

$$\mathbb{E}\left[\widehat{\text{Var}}^{\text{BC}}(\beta)\right] = \widehat{\text{Var}}(\beta).$$

Moreover, under suitable regularity conditions (which in particular require sufficient within-unit variation in X_{it}) we have, as n tends to infinity while T is kept fixed,

$$\widehat{\text{Var}}^{\text{BC}}(\beta) = \widehat{\text{Var}}(\beta) + o_p(1).$$

Note that, unlike $\widehat{\text{Var}}(\widehat{\beta})$, $\widehat{\text{Var}}^{\text{BC}}(\beta)$ may be negative in a given sample.

Exact bias-correction strategies can be extended to a number of other quantities in linear regression models such as (6). Arellano and Bonhomme (2012) show how to obtain consistent

estimators in short panels of higher-order moments of effects, such as skewness and kurtosis. They also show how to estimate the entire distribution of effects consistently, by adapting nonparametric deconvolution techniques (e.g., Stefanski and Carroll, 1990, Li and Vuong, 1998, Bonhomme and Robin, 2010).

A practical challenge for exact bias reduction methods is that they require correct modeling of the dependence of errors. If the dependence is misspecified, for example if the ε_{it}'s are assumed to be serially independent but are in fact serially correlated in the data, then the resulting estimators are no longer consistent. Moreover, exact bias correction is generally not available in nonlinear models, such as binary or multinomial choice models.

When the conditions for exact bias correction are not met, it is nevertheless often possible to achieve approximate bias correction, which becomes increasingly accurate as the panel length grows, and is expected to work well in panels of moderate length. We now describe several approximate bias-correction methods.

4.2 Approximate bias correction : the large-T perspective

The approximate bias-correction approach is motivated from a time-series perspective. Fixed-effects estimators such as $\widehat{\beta}_i$ are constructed based on T individual observations. When T is sufficiently large and serial dependence is not too strong, one can rely on large-T approximation arguments justified by central limit theorems to approximate the distribution of $\widehat{\beta}_i$. This perspective, which can be applied to general classes of models, including nonlinear and dynamic models, has been developed in a large literature, see Hahn and Kuersteiner (2002), Arellano (2003a), Hahn and Newey (2004), Arellano and Hahn (2007), and Fernández-Val and Weidner (2018), among others.

To illustrate the approach, consider a fixed-effects estimator $\widehat{\beta}_i$. As T tends to infinity, one often can write

$$\widehat{\beta}_i = \beta_i + \frac{1}{\sqrt{T}} Z_i + o_p\left(\frac{1}{\sqrt{T}}\right), \tag{13}$$

where Z_i is normally distributed with zero mean. The expansion (13) can be justified by appealing to a central limit theorem for serially dependent data. This requires that dependence is not too strong.

As an example, $\widehat{\beta}_i$ in model (8), which is equal to

$$\widehat{\beta}_i = \beta_i + \frac{\sum_{t=1}^{T} \left(X_{it} - \overline{X}_i\right) \varepsilon_{it}}{\sum_{t=1}^{T} \left(X_{it} - \overline{X}_i\right)^2},$$

can be written as (13) for $Z_i \sim \mathcal{N}(0, V_i)$, where V_i is the long-run variance

$$V_i = \operatorname*{plim}_{T\to\infty} \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[\left(X_{it}-\mu_i\right)\left(X_{is}-\mu_i\right)\varepsilon_{it}\varepsilon_{is}\right]}{\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(X_{it}-\mu_i\right)^2\right]\right\}^2}, \tag{14}$$

for $\mu_i = \operatorname{plim}_{T\to\infty} \overline{X}_i$. The assumption that V_i is finite limits the amount of serial dependence in ε_{it}.

In turn, an expansion of the form (13) typically implies that the bias of a fixed-effects estimator of a moment (such as the mean or variance of effects) or distribution takes the form

$$\text{Bias} = \frac{\text{B}}{T} + o\left(\frac{1}{T}\right), \tag{15}$$

where B is a constant.

A variety of methods have been developed to reduce bias based on expansions like (15). A key insight is that, if one can construct an estimator $\widehat{\text{B}}$ that is consistent for B as T tends to infinity, then subtracting $\frac{\widehat{\text{B}}}{T}$ from a fixed-effects estimator will deliver a bias-reduced estimator, in the sense that the resulting estimator will have a lower bias of order $o(1/T)$, instead of $O(1/T)$.

To illustrate, in the case of model (8) without additional covariates, we have

$$\mathbb{E}\left[\widehat{\text{Var}}\left(\widehat{\beta}\right)\right] = \widehat{\text{Var}}\left(\beta\right) + \underbrace{\frac{1}{T}\left(1-\frac{1}{n}\right)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}V_i\right] + o\left(\frac{1}{T}\right)}_{\text{Bias}}.$$

Using a consistent estimator of V_i as T tends to infinity, which we denote as $\widehat{V}_i$ (for example, a Newey-West estimator), we can then construct the bias-reduced estimator

$$\widehat{\text{Var}}^{\text{BR}}\left(\beta\right) = \widehat{\text{Var}}\left(\widehat{\beta}\right) - \frac{1}{T}\left(1-\frac{1}{n}\right)\frac{1}{n}\sum_{i=1}^{n}\widehat{V}_i,$$

which is guaranteed to satisfy

$$\mathbb{E}\left[\widehat{\text{Var}}^{\text{BR}}\left(\beta\right)\right] = \widehat{\text{Var}}\left(\beta\right) + o\left(\frac{1}{T}\right). \tag{16}$$

Exact and approximate bias-correction methods have both advantages and drawbacks. Compared to $\widehat{\text{Var}}^{\text{BC}}\left(\beta\right)$, which is unbiased irrespective of n and T, $\widehat{\text{Var}}^{\text{BR}}\left(\beta\right)$ is biased for fixed T. However, its bias is of a smaller order in T compared to that of the original fixed-effects estimator, see (16). On the other hand, the approximate bias reduction property, in the sense of (16), does not require correct specification of the serial dependence of ε_{it}, while exact bias correction does.

4.3 Half-panel jackknife

Among the available methods for approximate bias correction, we now describe a simple approach based on the jackknife proposed by Dhaene and Jochmans (2015). Let $\hat{\beta}^{(n,1:T/2)}$ be the $n \times 1$ vector of fixed effects estimated on the first half of the panel, i.e., only using periods $1, ..., T/2$ (taking the integer part if T is odd). Suppose in addition that observations are stationary over time. Then, the same logic that led to (15) implies that

$$\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\beta}^{(n,1:T/2)}\right)\right] = \widehat{\text{Var}}\left(\beta\right) + \frac{B}{T/2} + o\left(\frac{1}{T}\right).$$

Hence, the bias is approximately $2B/T$, which is twice as large as the bias of $\widehat{\text{Var}}\left(\hat{\beta}\right)$ based on the full sample.

This suggests constructing the *half-panel jackknife* estimator

$$\widehat{\text{Var}}^{\text{HPJ}}\left(\beta\right) = 2\widehat{\text{Var}}\left(\hat{\beta}\right) - \frac{1}{2}\left(\widehat{\text{Var}}\left(\hat{\beta}^{(n,1:T/2)}\right) + \widehat{\text{Var}}\left(\hat{\beta}^{(n,T/2+1:T)}\right)\right),$$

where $\hat{\beta}_i^{(n,T/2+1:T)}$ denote fixed-effects estimators based on the second half of the panel. Indeed, we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\text{Var}}^{\text{HPJ}}\left(\beta\right)\right] &= 2\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\beta}\right)\right] - \frac{1}{2}\left(\mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\beta}^{(n,1:T/2)}\right)\right] + \mathbb{E}\left[\widehat{\text{Var}}\left(\hat{\beta}^{(n,T/2+1:T)}\right)\right]\right) \\
&= 2\left(\widehat{\text{Var}}\left(\beta\right) + \frac{B}{T} + o\left(\frac{1}{T}\right)\right) - \frac{1}{2}\left(2\widehat{\text{Var}}\left(\beta\right) + 2\frac{B}{T/2} + o\left(\frac{1}{T}\right)\right) \\
&= \widehat{\text{Var}}\left(\beta\right) + o\left(\frac{1}{T}\right),
\end{aligned}$$

which implies that the bias of the half-panel jackknife estimator is of lower order compared to that of the fixed-effects estimator. A practical advantage of the jackknife approach is that there is no need to estimate the constant B in (15). In addition to this ease of implementation, half-panel jackknife is robust to the presence of serial correlation, in contrast with leave-one out jackknife (e.g., Hahn and Newey, 2004).

However, stationarity over time may be a restrictive assumption. For example, stationarity may fail in the presence of time fixed-effects. Fernández-Val and Weidner (2016) propose a modification of half-panel jackknife that handles the presence of time effects. To describe their approach, let $\hat{\beta}^{(1:n/2,T)}$ denote the $(n/2) \times 1$ vector of fixed effects estimated using the first half of individual units, with a similar notation for the other sub-sample. Then the modified half-panel jackknife estimator

$$\begin{aligned}
\widehat{\text{Var}}^{\text{HPJ2}}\left(\beta\right) = 3\widehat{\text{Var}}\left(\hat{\beta}\right) &- \frac{1}{2}\left(\widehat{\text{Var}}\left(\hat{\beta}^{(n,1:T/2)}\right) + \widehat{\text{Var}}\left(\hat{\beta}^{(n,T/2+1:T)}\right)\right) \\
&- \frac{1}{2}\left(\widehat{\text{Var}}\left(\hat{\beta}^{(1:n/2,T)}\right) + \widehat{\text{Var}}\left(\hat{\beta}^{(n/2+1:n,T)}\right)\right)
\end{aligned}$$

has reduced bias, even in the presence of time effects. Intuitively, time effects induce an additional bias of order $O(1/n)$, which the split in the cross-sectional dimension helps correct for.

There exist various alternatives to half-panel jackknife. Methods based on analytical corrections (which aim to find an empirical counterpart $\widehat{B}$ to B in (15)) have been developed for fixed-effects estimators, moment equations, and likelihood functions. Various bootstrap and jackknife methods have been proposed. Arellano and Hahn (2007) provides a comprehensive survey of those approaches. Hahn, Hughes, Kuersteiner, and Newey (2022) compare the efficiency of various approximate bias-correction methods.

Lastly, while we have used the variance of effects as an example to illustrate the methods throughout this section, approximate bias correction can be applied to other quantities. For example, Jochmans and Weidner (2024) consider the distribution function of effects

$$\widehat{F}_{\beta}(b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\beta_i \leqslant b\right\}.$$

They show that the distribution function of the fixed effects,

$$\widehat{F}_{\widehat{\beta}}(b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{\beta}_i \leqslant b\right\}$$

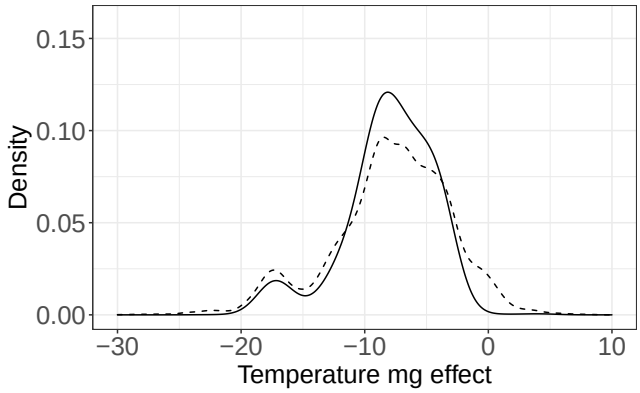
is biased for $\widehat{F}_{\beta}(b)$, with a bias of order $O(1/T)$ under suitable conditions. They then develop analytical and jackknife methods for approximate bias correction. In particular, they show that half-panel jackknife reduces bias.

4.4 Illustration

To bias-correct our empirical estimates, we apply the modified half-panel jackknife estimator HPJ2 that handles the presence of time effects (Dhaene and Jochmans, 2015, Fernández-Val and Weidner, 2016). Since the model includes state-year fixed effects, we implement the cross-sectional splits within each state. We use 5 splits, and average over them. In the time-series dimension we simply split the sample in half, only once. We apply the jackknife separately to the mean of the effects, their variance, and their distribution function.

In Column (3) of Table 2 we see that the average effect is virtually unaffected by the bias correction (-8.0, compared to -7.9). However, the variance is greatly reduced by the bias correction. The standard deviation decreases from 5.0 (uncorrected) to 3.6 (bias-corrected us-

Figure 5: Jackknife results – marginal density



Notes: density of marginal temperature effects across counties and years, weighted by corn area. The solid line shows the density according to jackknife, while the dotted line shows the density according to fixed effects.

Table 2: Distribution of marginal effects of temperature across counties and years

	Homog.	Heterog.	Jackknife	GFE	RE
	(1)	(2)	(3)	(4)	(5)
Percentile 10	-6.600	-14.701	-12.510	-16.004	-13.745
Percentile 25	-6.600	-10.255	-9.477	-10.731	-10.452
Percentile 50	-6.600	-7.416	-7.902	-7.409	-7.625
Percentile 75	-6.600	-4.574	-5.283	-4.175	-5.025
Percentile 90	-6.600	-2.356	-3.789	-4.175	-2.826
Mean	-6.600	-7.856	-8.042	-7.796	-7.950
Variance	0.000	24.850	12.621	16.085	18.221

Notes: distribution of marginal effects across counties and years, weighted by corn area. Column (1) corresponds to a homogeneous β. Columns (2) to (5) correspond to heterogeneous β_i: (2) is based on fixed-effects, (3) on the jackknife, (4) on grouped fixed-effects with $K = 5$, and (5) on correlated random-effects with a prior mean that depends on temperature.

ing the jackknife).[9] We also report estimates of percentiles of the distribution of β_i. Those are obtained by inverting the jackknife estimate of the distribution function (Jochmans and

[9]In Table 2 we show the mean and variance estimates implied by the bias-corrected distribution. Bias-correcting the mean directly gives a similar value (-8.1). However, bias-correcting the variance directly gives an even lower dispersion, with a standard deviation of 2.7.

Weidner, 2024). In practice, we find that the estimated distribution function is not monotone everywhere, so we apply the rearrangement method of Chernozhukov, Fernández-Val, and Galichon (2010) to enforce monotonicity. The estimates of the percentiles in Column (3) of Table 2 are consistent with the finding that the jackknife tends to compress the distribution of effects. This is further corroborated by Figure 5, where we report a kernel-smoothed estimate of the bias-corrected density.[10]

5 Discrete heterogeneity: grouping methods

The noise in the fixed-effects estimates $\widehat{\beta}_i$ reflects the fact that there is not enough data to precisely estimate all β_i's. While the sample contains nT observations, only T of them are informative about each β_i. To reduce the impact of noise, a common approach is to impose additional assumptions on the heterogeneity. Such strategies are commonly employed in high-dimensional models. As an example, a *sparse* model assumes that all β_i's are equal to zero (or to a common coefficient), except for a few of them that are non-zero (e.g., Tibshirani, 1996). Another example assumes that the β_i's follow a Gaussian distribution, and we will review such *random-effects* approaches in Section 6.

In this section we describe another assumption about the β_i's, which has empirical appeal for modeling heterogeneity. The idea is to limit the number of different values that β_i can take. While unrestricted β_i's can take up to n different values, a *discrete heterogeneity* assumption sets the maximum number of distinct values to K, where K is typically much smaller than n.

5.1 Model and estimator

Suppose that individual units are partitioned into K groups, $k_i \in \{1, ..., K\}$, and that, within group k, all β_i's are constant equal to $\underline{\beta}_k$. We then consider the following grouped fixed-effects model

$$Y_{it} = \underline{\beta}_{k_i} X_{it} + \alpha_i + \delta_t + W_{it}'\gamma + \varepsilon_{it}. \tag{17}$$

When $K = n$, (17) simplifies to the fixed-effects model (6). However, taking K much smaller than n pools information across individual units. This can help reduce noise, and thus alleviate

[10] To assess whether these findings are affected by low variation in temperature over time for some counties, in Appendix Figure B8 we report the results of a trimming exercise where we remove some share of counties that exhibit the lowest variation. The estimates are quite stable as a function of the share of observations removed.

the first three issues with fixed-effects estimation that we mentioned in Section 3.[11]

There are several available methods to estimate model (17). One strategy is to compute a grouped fixed-effects estimator, by minimizing, for a given number of groups K, the objective function

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it}-\underline{\beta}_{k_i}X_{it}-\alpha_i-\delta_t-W_{it}'\gamma\right)^2 \tag{18}$$

with respect to the following parameters: the unit fixed-effects intercepts $\alpha_1,...,\alpha_n$, the time fixed-effects intercepts $\delta_1,...,\delta_T$, the covariates' coefficients γ, the group-specific parameters $\underline{\beta}_1,...,\underline{\beta}_K$, and the unit-specific group membership indicators $k_1,...,k_n$. The group indicators $k_i \in \{1,...,K\}$ define a partition of all individual units. The minimum in (18) is taken over all possible such partitions into K groups.

The grouped fixed effects (or GFE) estimator defined as the minimizer of (18) can be interpreted as a generalization of *kmeans clustering*, which is a widely used partitioning algorithm. In fact, one could alternatively estimate the groups by applying kmeans to the fixed-effects $\widehat{\beta}_i$, that is, by minimizing

$$\sum_{i=1}^{n}\left(\widehat{\beta}_i-\underline{\beta}_{k_i}\right)^2, \tag{19}$$

with respect to the group-specific parameters $\underline{\beta}_1,...,\underline{\beta}_K$ and the unit-specific group membership indicators $k_1,...,k_n$. An advantage of (18) compared to (19) is that the estimator remains well-defined even if some of the $\widehat{\beta}_i$'s do not exist. In other words, by minimizing (18) one can address the second issue with fixed-effects estimation, which is the non-existence of fixed-effects estimates. Heterogeneous estimators based on (18) are sometimes referred to as "clusterwise regression" estimators in computer science.

There is an extensive literature in statistics and computer science that proposes and studies algorithms for kmeans clustering and clusterwise regression. A particularly simple approach is based on Lloyd's algorithm. The method consists in iterating between a grouping step and a regression step, as follows:

- In the *grouping step*, given some values of the coefficients $\alpha,\delta,\gamma,\underline{\beta}$, one minimizes (18) with respect to $k_1,..,k_n$. The solution is

$$k_i=\underset{k=1,...,K}{\operatorname{argmin}}\sum_{t=1}^{T}\left(Y_{it}-\underline{\beta}_{k}X_{it}-\alpha_i-\delta_t-W_{it}'\gamma\right)^2, \quad \text{for all } i=1,...,n.$$

[11] This is not yet addressing the fourth issue, since here we maintain the assumption that β_i is time-invariant. However, in Section 7 we will review grouped-factor models that allow for time-varying heterogeneity.

- In the *regression step*, given $k_1, ..., k_n$ one minimizes

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \underline{\beta}_{k_i} X_{it} - \alpha_i - \delta_t - W_{it}'\gamma\right)^2$$

  with respect to all parameters (except for the group indicators k_i that are fixed) and obtain $\alpha', \delta', \gamma', \underline{\beta}'$. This step is simply a linear regression, where some covariates are interactions between X_{it} and group indicators. One then sets $(\alpha, \delta, \gamma, \underline{\beta}) = (\alpha', \delta', \gamma', \underline{\beta}')$, and go back to the previous step, iterating until the objective function does not change.

In practice, the objective function tends to have many local minima, and it is important to start the algorithm from multiple parameter values. We will illustrate this approach in the application. The choice of initialization has been studied in computer science, see for example Arthur and Vassilvitskii (2007). Moreover, Lloyd's algorithm is only one simple approach to minimize (18). There exists a variety of alternatives, both approximate and exact, some of which are reviewed in Bonhomme and Manresa (2015).

In addition to grouped fixed-effects estimation, other methods have been proposed to estimate group membership indicators and regression parameters in discrete heterogeneity models such as (17). The Classifier Lasso (or CLasso) approach proposed by Su, Shi, and Phillips (2016) is a penalized regression method based on minimizing

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \beta_i X_{it} - \alpha_i - \delta_t - X_{it}'\gamma\right)^2 + \lambda \sum_{i=1}^{n}\prod_{k=1}^{K}\left|\beta_i - \underline{\beta}_k\right| \tag{20}$$

with respect to the following parameters: the unit fixed-effects intercepts $\alpha_1, ..., \alpha_n$, the time fixed-effects intercepts $\delta_1, ..., \delta_T$, the covariates' coefficients γ, the unit fixed-effects coefficients $\beta_1, ..., \beta_n$, and the group-specific parameters $\underline{\beta}_1, ..., \underline{\beta}_K$.

In (20), $\lambda > 0$ is a tuning parameter. Classifier Lasso exhibits a behavior that is reminiscent of the Lasso. Indeed, the presence of the absolute value $|\beta_i - \underline{\beta}_k|$ in the penalty term tends to produce a clustering of the β_i's around the group-specific values $\underline{\beta}_1, ..., \underline{\beta}_K$. This is reminiscent of the Lasso penalty, which tends to produce a clustering of estimates around zero. However, unlike the Lasso objective, (20) is not convex, and it can have local minima.

The literature on grouping methods is evolving fast. Recent approaches include Chetverikov and Manresa (2022), Mugnier (2022), Yu, Gu, and Volgushev (2024), and Yu, Gu, and Volgushev (2022), among others. These methods, and the Classifier Lasso, all share similar asymptotic guarantees, which we will review next.

In practice, the number of groups K is an important input to grouping methods. Several methods have been proposed to estimate K. One approach is based on information criteria.

Let $\widehat{L}(K)$ denote the value of the objective function in (18) at the estimated parameters, for a given number of groups K. Su, Shi, and Phillips (2016) propose to minimize the Information Criterion

$$\log\left(\frac{1}{nT}\widehat{L}(K)\right) + \rho_{nT}K, \tag{21}$$

where ρ_{nT} is a tuning parameter, which they recommend to set as $\rho_{nT} = \frac{2}{3}(nT)^{-\frac{1}{2}}$. Bonhomme and Manresa (2015) consider the Bayesian Information Criterion

$$\frac{1}{nT}\widehat{L}(K) + \frac{K}{nT}\ln(nT)\widehat{\sigma}^2, \tag{22}$$

where $\widehat{\sigma}^2 = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(Y_{it} - \widehat{\beta}_i X_{it} - \widehat{\alpha}_i - \widehat{\delta}_t - X_{it}'\widehat{\gamma}\right)^2$ is based on the fixed-effects estimator of model (6). Another approach to estimate K is through sequential testing. Lu and Su (2017) propose an LM test statistic to test the null hypothesis $K = k$ against $K > k$, for $k = 1, 2, ...$ The estimator of K is then the first value of k for which the null hypothesis is not rejected.

5.2 Properties under grouped heterogeneity

There are two approaches to provide theoretical justifications for grouping methods. The *first* and most common approach is to analyze the behavior of grouped heterogeneity estimates in a data generating process where the heterogeneity is indeed grouped in the population. In this approach, the DGP is assumed to satisfy model (17), for some true number of groups K and true group membership indicators $k_1, ..., k_n$. In particular, one assumes that there exists a true number of groups K that does not depend on the sample size.

In a grouped data generating process, several papers have provided conditions for the groups to be consistent as n and T tend to infinity. See Hahn and Moon (2010), Lin and Ng (2012), Bonhomme and Manresa (2015), and Su, Shi, and Phillips (2016), among others. This implies that the probability of correctly classifying all individual units tends to one as the sample size tends to infinity. Formally, group consistency reads[12]

$$\Pr\left[\widehat{k}_i = k_i \text{ for all } i = 1, ..., n\right] \to 1. \tag{23}$$

Conditions for group consistency allow n to grow polynomially faster than T. For example, Bonhomme and Manresa (2015) assume that $n/T^{\delta} \to 0$ for some $\delta > 0$. Since δ can be

[12]Since the groups are not observed, the definition of groups $1, 2, 3, ...$ is arbitrary. Indeed, one could alternatively refer to group 1 as group 2, and to group 2 as group 1, without changing model (17). Group consistency in (23) is thus understood to hold for an arbitrary labeling of the groups.

arbitrarily small, this theory provides a rationale for using grouping estimators in panel data where the time dimension is much smaller than the cross-sectional dimension.

An implication of group consistency in (23) is that the asymptotic distribution of grouped estimates of heterogeneous effects is not affected by the fact that the groups have been estimated. Consider the estimator $\widetilde{\underline{\beta}}_k$ of $\underline{\beta}_k$ based on the *true groups* k_i. It follows from group consistency that the grouped fixed-effects estimator of β_i, which we have denoted as $\widehat{\underline{\beta}}_{\widehat{k}_i}$, has the same asymptotic distribution as $\widetilde{\underline{\beta}}_{k_i}$. In particular, when all groups have non-negligible size in the limit, $\widehat{\underline{\beta}}_{\widehat{k}_i}$ is $\sqrt{nT}$-consistent for β_i. This should be contrasted to the fixed-effects estimator $\widehat{\beta}_i$, which is only $\sqrt{T}$-consistent. This provides a concrete sense in which grouping can reduce noise and improve the performance of heterogeneity estimates.

As a result of group consistency, inference on the parameters of the grouped fixed-effects model (17) is very simple. One can proceed as if the groups k_i were known to the researcher, and simply replace k_i by their estimates $\widehat{k}_i$ in formulas for standard errors and confidence intervals. Moreover, under related conditions, the number of groups can be shown to be consistently estimated using information criteria or sequential testing.

However, group consistency hinges on several possibly restrictive assumptions. Bonhomme and Manresa (2015) emphasize three conditions. The first one is that the true groups need to be sufficiently well separated. The theory assumes that the distance $|\overline{\beta}_k - \overline{\beta}_{k'}|$, for any groups $k \neq k'$, is non-zero and fixed as the sample size tends to infinity. This may not well capture the fact that, in a fixed sample, some groups may be close to each other, and hence hard to distinguish. The second and third conditions require errors ε_{it} to be weakly dependent and their distributions to have sufficiently thin tails.

There is relatively little work trying to provide valid inference methods when group consistency does not hold. Recently, Armstrong, Weidner, and Zeleneev (2022) studied the related problem of inference in interactive fixed-effects models in the presence of weak factors, and provided bias-aware inference methods in such settings. Developing inference methods that correctly account for the uncertainty in group assignments is an important avenue for future work.[13]

[13] In their supplement, Bonhomme and Manresa (2015) used the bootstrap, clustered at the unit level, in an attempt to account for the uncertainty in group estimation. In a recent contribution, Dzemski and Okui (2024) proposed a method for confidence intervals for group membership indicators.

5.3 Properties under continuous heterogeneity

The *second* approach to analyze the properties of grouped heterogeneity estimators is to assume a data generating process where heterogeneity is continuous. In this approach, the DGP is assumed to satisfy (6), where one does not assume that the β_i's are grouped. This reflects the view that groups provide an approximation to some possibly continuous heterogeneity.

In pioneering contributions in the early 1980s, David Pollard showed that, when heterogeneity β_i is continuous, group estimates converge to some *pseudo-true* values as n tends to infinity for T fixed (Pollard, 1981, Pollard, 1982). While Pollard considered the case of kmeans, his results apply more generally to grouped fixed-effects estimates (Bonhomme and Manresa, 2015). However, these pseudo-true values differ from the true parameter values, and grouped fixed-effects estimates are inconsistent as n tends to infinity if T is fixed.

Recently, Bonhomme, Lamadon, and Manresa (2022) showed that, in settings with continuous β_i's, parameter estimates remain consistent for their true values as n and T tend to infinity jointly. However, a crucial difference when the β_i's are not discrete is that consistency only holds as K tends to infinity together with the sample size. A too small number of groups provides a poor approximation to the heterogeneity, which in turn affects consistency and convergence rates. When heterogeneity is continuous, the groups are a regularization device, with K being a tuning parameter. Moreover, the convergence rate of grouped fixed-effects estimators depends crucially on the dimensionality of heterogeneity. From this perspective, the case of model (6) appears favorable since heterogeneity β_i is scalar. However, the performance of grouping methods may worsen in models with multi-dimensional continuous heterogeneity.[14]

5.4 Illustration

We implement the grouped fixed-effects estimator in our illustration, for various numbers of groups. For computation we use Lloyd's algorithm with 3,000 starting values. In Appendix C we describe how we initialize the algorithm. Based on the estimates for K groups, where K ranges between 1 and 8, we compute the information criterion (21) proposed by Su, Shi, and Phillips (2016). According to this criterion, the optimal number of groups is 5. Given this, we

[14] Cheng, Schorfheide, and Shao (2023) propose a method to allow for separate group indicators in the parameters of a regression model. For example, with X_{it} denoting temperature and W_{it} denoting precipitations, one can specify

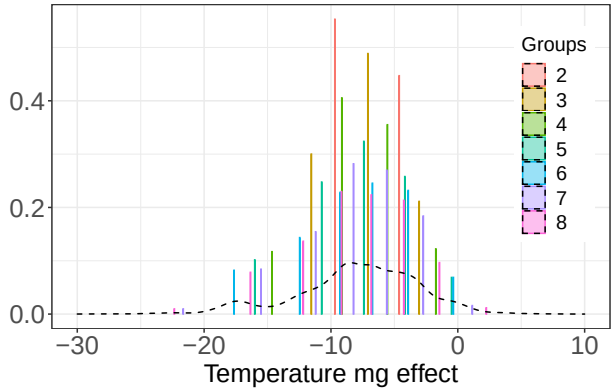
$$Y_{it} = \underline{\beta}_{k_i} X_{it} + \underline{\gamma}_{\ell_i} W_{it} + \alpha_i + \delta_t + \varepsilon_{it},$$

where $k_i \in \{1, ..., K\}$ and $\ell_i \in \{1, ..., L\}$. Here there are two numbers of groups K, L, and two group indicators k_i, ℓ_i. In their asymptotic analysis they consider DGPs with discrete heterogeneity.

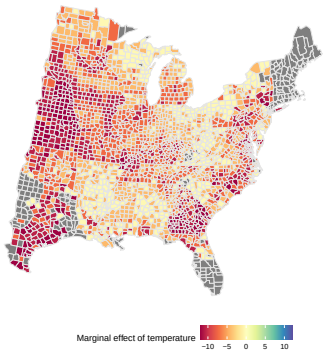
set $K = 5$ as a baseline. However, we report results based on varying numbers of groups in Appendix Table B5 and Appendix Figure B9. The results are quite stable between $K = 3$ and $K = 8$.

Figure 6: Grouped fixed-effects estimates

(a) Histograms



(b) Spatial distribution ($K = 5$)



Notes: panel (a) shows histograms of marginal effects across counties and years, weighted by corn area, based on grouped fixed-effects for several values of K, while the dotted line shows the density of fixed-effects estimates. Panel (b) shows the spatial distribution of marginal effects across counties based on grouped fixed-effects with $K = 5$.

In Column (4) in Table 2 we report the mean, variance, and percentiles of the distribution of temperature impacts according to grouped fixed-effects, for $K = 5$. The mean is again very similar to the uncorrected fixed-effects average (-7.8). The dispersion of effects is lower than the one of fixed-effects estimates, but larger than the jackknifed one, with a standard deviation of 4.0. As shown by the histograms in panel (a) of Figure 6, plotted for various numbers of groups, a difference between the uncorrected fixed-effects density and the grouped fixed-effects histograms is the absence of counties with *positive* temperature impacts in the latter case when $K \leqslant 6$.

Grouped fixed-effects estimates can be used to represent spatial heterogeneity. In panel (b) of Figure 6 we report the grouped fixed-effects estimates $\underline{\widehat{\beta}}_{\widehat{k}_i}$ on a map of the US, for $K = 5$. Under a model with discrete heterogeneity, and suitable assumptions as we have discussed, these estimates are $\sqrt{nT}$-consistent, whereas the fixed-effects estimates are only $\sqrt{T}$-consistent. This provides a rationale for reporting, and interpreting, grouped estimates such as the ones shown in panel (b) of Figure 6 in applications.

6 Modeling heterogeneity: random-effects methods

6.1 Random-effects model

Another approach to try to reduce the impact of the noise on heterogeneity estimates is to model the *distribution* of heterogeneous parameters β_i. This approach is increasingly used in economics, in panel data but also in settings with a network structure such as in the estimation of neighborhood effects or firm and worker effects (see Bonhomme and Denis, 2024).

To illustrate this approach in our context, consider the fixed-effects estimates

$$\widehat{\beta}_i = \beta_i + v_i, \tag{24}$$

where by (13) $\sqrt{T}v_i = Z_i + o_p(1)$, for $Z_i \,|\, V_i \sim \mathcal{N}(0, V_i)$. Asymptotic normality of v_i as T tends to infinity holds under suitable conditions on time-series dependence. For the presentation we will assume that normality holds exactly, and not only asymptotically, so

$$v_i \,|\, V_i \sim \mathcal{N}\left(0, \frac{V_i}{T}\right). \tag{25}$$

In addition, we will assume that V_i is known (see (14)), although in practice it needs to be estimated. Note that by Assumption 3 all observations are independent. Cross-sectional independence is a substantive assumption for the methods reviewed in this section.

Instead of viewing β_i as a parameter to be estimated (the so-called "fixed-effects" approach), suppose that it is drawn from some distribution G conditional on the variance V_i (the so-called "random-effects" approach). That is, suppose that

$$\beta_i \,|\, V_i \sim G(\cdot \,|\, V_i). \tag{26}$$

As an example, suppose that

$$\beta_i \,|\, V_i \sim \mathcal{N}\left(\mu_0 + \mu_1 V_i, \tau^2\right), \tag{27}$$

for some parameters μ_0, μ_1, τ^2. This so-called"correlated random-effects" specification allows the distribution of β_i to depend on V_i through a parametric model. In practice, one can condition on additional covariates, and we will follow this approach in the illustration by conditioning on average temperature in the county.

6.2 Estimators and properties

By combining (24) and (27), we have

$$\mathbb{E}\left[\widehat{\beta}_i \,|\, V_i\right] = \mu_0 + \mu_1 V_i, \tag{28}$$

$$\mathbb{E}\left[\left(\widehat{\beta}_i - \mu_0 - \mu_1 V_i\right)^2 \,|\, V_i\right] = \tau^2 + \frac{V_i}{T}. \tag{29}$$

This suggests to estimate $\widehat{\mu}_0, \widehat{\mu}_1$ by an OLS regression of $\widehat{\beta}_i$ on V_i and a constant, and to estimate

$$\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\beta}_i - \widehat{\mu}_0 - \widehat{\mu}_1 V_i\right)^2 - \frac{1}{nT}\sum_{i=1}^{n} V_i. \tag{30}$$

The term $\frac{1}{nT}\sum_{i=1}^{n} V_i$ can be interpreted as an exact bias correction when estimating the conditional variance τ^2 of the effects β_i. The unconditional variance can then be estimated as

$$\widehat{\mathrm{Var}}^{\mathrm{RE}}(\beta) = \widehat{\mathrm{Var}}\left(\widehat{\mu}_0 + \widehat{\mu}_1 V_i\right) + \widehat{\tau}^2.$$

The estimators $\widehat{\mu}_0$, $\widehat{\mu}_1$, $\widehat{\tau}^2$, and $\widehat{\mathrm{Var}}^{\mathrm{RE}}(\beta)$ are all consistent as n tends to infinity and T fixed under standard regularity conditions, provided (27) holds.

An attractive feature of the random-effects approach is the ability to compute posterior quantities. Indeed, interpreting (27) as a prior for β_i (conditional on V_i) in model (24)-(25), we obtain the posterior distribution of β_i given the data, which is

$$\beta_i \,|\, \widehat{\beta}_i, V_i \sim \mathcal{N}\left(\left(\frac{\tau^2}{\frac{V_i}{T} + \tau^2}\right)\widehat{\beta}_i + \left(\frac{\frac{V_i}{T}}{\frac{V_i}{T} + \tau^2}\right)\left(\mu_0 + \mu_1 V_i\right), \frac{1}{\left(\frac{V_i}{T}\right)^{-1} + \tau^{-2}}\right), \tag{31}$$

independent across i. In particular, the posterior mean of β_i is

$$\mathbb{E}\left[\beta_i \,|\, \widehat{\beta}_i, V_i\right] = \left(\frac{\tau^2}{\frac{V_i}{T} + \tau^2}\right)\widehat{\beta}_i + \left(\frac{\frac{V_i}{T}}{\frac{V_i}{T} + \tau^2}\right)\left(\mu_0 + \mu_1 V_i\right). \tag{32}$$

It is the best predictor of β_i under quadratic loss in model (24)-(25)-(27).

Let us introduce the shrinkage factor

$$\rho_i = \frac{\tau^2}{\frac{V_i}{T} + \tau^2},$$

which lies between 0 and 1. In (32), the fixed-effects estimate $\widehat{\beta}_i$ is shrunk towards the prior mean. The shrinkage is more aggressive when ρ_i is lower, which corresponds to a lower signal-to-noise ratio.

Given estimates $\widehat{\mu}_0$, $\widehat{\mu}_1$, and $\widehat{\tau}^2$, one can construct the empirical-Bayes posterior means

$$\widehat{\beta}_i^{\text{PM}} = \left(\frac{\widehat{\tau}^2}{\frac{V_i}{T} + \widehat{\tau}^2} \right) \widehat{\beta}_i + \left(\frac{\frac{V_i}{T}}{\frac{V_i}{T} + \widehat{\tau}^2} \right) (\widehat{\mu}_0 + \widehat{\mu}_1 V_i) \,. \tag{33}$$

Shrunk estimates (33) have attractive properties as square loss minimizers, even in cases where (27) is misspecified (see the James-Stein theorem). Intuitively, shrinkage helps because using information from other individuals $i' \neq i$ can improve prediction for individual i (Efron, 2012, Koenker and Gu, in preparation).

Often, researchers are interested in quantities averaged over individual units, such as the dispersion of effects or their distribution. Given the random-effects model (24)-(25)-(27), one can report posterior mean estimates of those quantities. For example, the posterior mean of the sample variance of effects is

$$\mathbb{E}\left[\widehat{\text{Var}}\left(\beta\right) \mid \widehat{\beta}_1, ..., \widehat{\beta}_n, V_1, ..., V_n \right] = \widehat{\text{Var}}\left(m\right) + \left(1 - \frac{1}{n}\right) \frac{1}{n} \sum_{i=1}^{n} s_i^2, \tag{34}$$

where m_i and s_i^2 are the posterior means and variances of β_i in (31). In turn, the posterior mean of the distribution of effects is

$$\mathbb{E}\left[\widehat{F}_\beta(b) \mid \widehat{\beta}_1, ..., \widehat{\beta}_n, V_1, ..., V_n \right] = \frac{1}{n} \sum_{i=1}^{n} \Phi\left(\frac{b - m_i}{s_i} \right), \tag{35}$$

where Φ denotes the standard normal distribution function.

Posterior estimates such as (34) and (35) have attractive robustness properties as T tends to infinity, as highlighted by Arellano and Bonhomme (2009) who refer to those as "Bayesian fixed-effects" estimates. As T tends to infinity,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\text{Var}}\left(\beta\right) \mid \widehat{\beta}_1, ..., \widehat{\beta}_n, V_1, ..., V_n \right] &= \widehat{\text{Var}}\left(\beta\right) + o_p(1) \\
\mathbb{E}\left[\widehat{F}_\beta(b) \mid \widehat{\beta}_1, ..., \widehat{\beta}_n, V_1, ..., V_n \right] &= \widehat{F}_\beta(b) + o_p(1),
\end{aligned}$$

even when the parametric model for β_i (e.g., (27)) is incorrectly specified (see also Hahn, Kuersteiner, and Cho, 2004). That is, consistency holds even when β_i are not normally distributed, or they are normal but their variance depends on V_i, for example. Moreover, Bonhomme and Weidner (2022) show that posterior estimates such as (34) and (35) also enjoy robustness properties for T fixed as n tends to infinity, even though misspecified random-effects estimators are generally inconsistent for fixed T.

Extension 1. The normal specification for β_i in (27) can be generalized. A fully nonparametric approach, independent of the conditioning variable V_i, was introduced by Kiefer and Wolfowitz (1956). Nonparametric maximum likelihood estimation is the subject of a large literature in statistics, see for example Koenker and Mizera (2014) and Gu and Koenker (2017). Flexible parametric methods have also been proposed, notably Efron (2016)'s penalized log-spline estimator. In applications of random-effects methods, it is often important to allow for dependence on conditioning variables such as V_i. Recently, Chen (2023) proposed an approach based on a semi-parametric location-scale model.

Extension 2. Our presentation of random-effects methods in this section has been based on model (24)-(25). This presupposes that $\widehat{\beta}_i$ has been computed for all i. However, as we pointed out as one of the main issues with fixed-effects, in practice it is often the case that some of the $\widehat{\beta}_i$'s cannot be calculated, for example due to insufficient variation in X_{it} over time. In such cases, an alternative approach is to work directly with the underlying regression model (6), and to specify a random-effects specification for β_i as, for example,

$$\beta_i \,|\, X_i, W_i, V_i \sim \mathcal{N}\left(\mu_0 + \mu_1 V_i, \tau^2\right). \tag{36}$$

Posterior estimators can then be computed under distributional assumptions on ε_{it} in (6), such as normality.

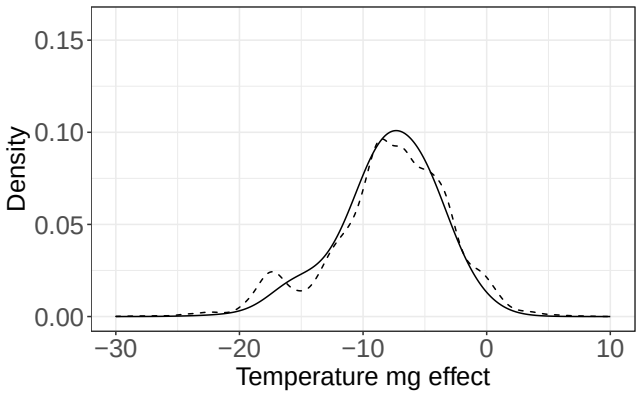
6.3 Illustration

We estimate random-effects specifications based on (24)-(25), for various priors. Our main specification is in the spirit of (27): β_i is assumed to be normally distributed, with a mean that depends linearly on average temperature in the county, and a conditional variance that is constant. In Column (5) of Table 2 we report posterior means of the average effect, the variance of effects, and the distribution of effects, which we invert to obtain percentiles. Again, the average is very similar to the other specifications (-8.0). The dispersion is slightly larger than the one under grouped fixed-effects, but lower than the variance of fixed-effects estimates, with a standard deviation of 4.3. In panel (a) of Figure 7 we report, in solid line, a smoothed kernel estimate of the density implied by the random-effects specification. We see relatively small differences with the estimated density of fixed effects in this case, except for somewhat thinner left and right tails. Lastly, in panel (b) of Figure 7 we plot estimates of posterior means of β_i on a map of the US.

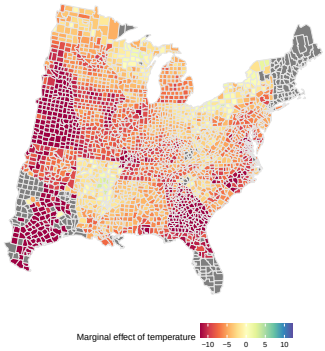
We perform several robustness checks. In Appendix Table B6 we report estimates based on different priors: a normal prior fully independent of covariates, and two additional priors that allow the conditional variance of β_i, in addition to its mean, to depend on average temperature in the county. The results are not very different across all these specifications. Another important practical question when implementing random-effects is how to estimate the variance of fixed-effects estimates, V_i. In our baseline, we use a within-county formula based on i.i.d. homoskedastic standard observations, but we have conducted several robustness checks using Newey-West estimators with various number of lags and found very similar random-effects estimates in those cases.

Figure 7: Random-effects estimates

(a) Marginal density



(b) Spatial distribution of posterior means



Notes: results weighted by corn area. Panel (a) presents the posterior mean of the distribution of marginal effects, kernel-smoothed to obtain a density (in solid), whereas the fixed-effects density is shown in dashed. Panel (b) maps the posterior means. The model assumes a normal prior whose mean depends on average temperature in the county.

7 Variation over time: factor methods

7.1 Low-rank models

Allowing for effects to vary over time is important in many applications. While a fully unrestricted β_{it} is not estimable, panel data, together with some modeling choices, offer the possibility to estimate models with time-varying, heterogeneous responses.

To illustrate, consider a simple departure from model (6), which allows for (a specific form

of) variation over time while maintaining essentially the same structure. Suppose that

$$\beta_{it} = \beta_i + \mu_t \tag{37}$$

is additive across units and time periods, and that

$$\begin{aligned} Y_{it} &= (\beta_i + \mu_t) X_{it} + \alpha_i + \delta_t + W_{it}'\gamma + \varepsilon_{it} \\ &= \beta_i X_{it} + \alpha_i + \delta_t + \mu_t X_{it} + W_{it}'\gamma + \varepsilon_{it}, \end{aligned}$$

which is identical to (6) except for the fact that the covariates now include interactions between X's and time indicators.

A key feature of (37) is that β_{it} depends on two sources of heterogeneity. The first one, β_i, captures unit heterogeneity. The second one, μ_t, captures time heterogeneity. This implies that the $n \times T$ matrix $\boldsymbol{B}$ with elements β_{it} in (7) can be written as a sum of a matrix with constant rows and another matrix with constant columns. In particular, under (37) $\boldsymbol{B}$ has rank 2. Such low-rank constraints are useful ways of disciplining the nature of heterogeneity across units and over time.

More generally, a factor model imposes a low-rank assumption on $\boldsymbol{B}$. A general representation of a rank-R matrix is

$$\beta_{it} = \sum_{r=1}^{R} \beta_{i,r}\mu_{r,t}. \tag{38}$$

Here, $\beta_{i,1}, ..., \beta_{i,R}$ are R sources of unit heterogeneity, while $\mu_{1,t}, ..., \mu_{R,t}$ represent R sources of time heterogeneity. In matrix form, $\boldsymbol{B}$ can be represented as a product

$$B = \boldsymbol{\beta}\boldsymbol{\mu}',$$

where $\boldsymbol{\beta}$ is an $n \times R$ matrix with elements $\beta_{i,r}$, and $\boldsymbol{\mu}$ is a $T \times R$ matrix with elements $\mu_{r,t}$.

Note that the usual parallel trends assumption in two-way fixed-effects,

$$\alpha_{it} = \alpha_i + \delta_t$$

takes the same form as (37). The specification for α_{it} can thus be generalized in exactly the same way to factor models with additional sources of possibly non-additive heterogeneity, as in (38). Hence, factor models also offer the possibility to relax the parallel trends assumption in two-way fixed-effects and other panel data regression settings. In fact, factor methods were initially applied to models with constant β, as we review next.

7.2 Factor methods

A well-studied model with a factor structure is the interactive fixed-effects model (Bai, 2009, Pesaran, 2006),

$$Y_{it} = \beta X_{it} + \alpha_i' \delta_t + \varepsilon_{it}, \tag{39}$$

where we do not impose parallel trends (and abstract from additional covariates for simplicity). Here α_i and δ_t are $R \times 1$ vectors, where R is the number of factors. Importantly, this model has constant β so it does not allow for heterogeneity across units or over time.

In model (39), Bai (2009) provides conditions under which the parameters β, $\alpha_1, ..., \alpha_n$ and $\delta_1, ..., \delta_T$ are all consistently estimated as n and T tend to infinity by minimizing[15]

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \left(Y_{it} - \beta X_{it} - \alpha_i{}' \delta_t\right)^2. \tag{40}$$

The resulting interactive fixed-effects estimator can be interpreted as principal component analysis (PCA) with covariates, and it is related to synthetic control (Abadie, Diamond, and Hainmueller, 2010, Gobillon and Magnac, 2016). However, the multiplicative structure $\alpha_i{}' \delta_t$ in (40) leads to a non-convex objective function, which complicates implementation.

An alternative approach is to enforce the low-rank constraint through a penalty, in the spirit of matrix completion methods. To proceed, let A denote an $n \times T$ matrix with elements α_{it}. Let $\|A\|_*$ denote the sum of the singular values of A.[16] This quantity is also called the nuclear norm of A. Moon and Weidner (2018) propose to minimize

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \left(Y_{it} - \beta X_{it} - \alpha_{it}\right)^2 + \lambda \|A\|_*, \tag{41}$$

with respect to β and the α_{it}'s, where the penalty parameter $\lambda > 0$, suitably chosen, amounts to restricting the rank of A. Indeed, the nuclear norm penalty can be understood as a matrix counterpart to the ℓ^1 norm that is used in Lasso estimation (Cai, Candès, and Shen, 2010, Hastie, Tibshirani, and Wainwright, 2015). Unlike (40), the objective in (41) is convex in A and β.

The use of nuclear norm penalization permits to extend factor methods to models where both α_{it} and β_{it} are heterogeneous, hence allowing for effects heterogeneity across units and over time. A recent contribution by Chernozhukov, Hansen, Liao, and Zhu (2019) proposed to extend (41) to also allow for coefficient heterogeneity.

[15] Identification of $\alpha_1, ..., \alpha_n$ and $\delta_1, ..., \delta_T$ requires choosing a suitable normalization, referred to as a choice of a "rotation".

[16] The singular values of A are the square roots of the eigenvalues of $A'A$.

7.3 Grouped-factor methods

However, while factor models address the fourth issue with fixed effects, by allowing for time variation, they are vulnerable to the other issues that we pointed out. Indeed, compared to a standard fixed-effects approach with time-invariant coefficients β_i, a factor specification such as (38) adds several (possibly many) parameters to estimate. However, as we have seen in the first part of the paper, one may want to *reduce* the number of fixed effects in practice because of sample noise.

To make progress, one can impose additional assumptions on β_{it} beyond the factor structure. In a *grouped-factor* model, we assume that

$$\beta_{it} = \underline{\beta}_t(k_i), \tag{42}$$

where $k_i \in \{1, ..., K\}$ are group membership indicators, and $\underline{\beta}_t(1)$, ..., $\underline{\beta}_t(K)$ are K group-specific paths of heterogeneity. Bonhomme and Manresa (2015) proposed this idea to model the intercept in a regression with constant coefficients. Here we advocate this approach to model the coefficients themselves.

The grouped-factor model (42) allows for effects heterogeneity across units and over time. In fact, this is a special case of a factor model with K factors, since we can write

$$\beta_t(k_i) = \sum_{r=1}^{K} \mathbf{1}\{k_i = r\}\underline{\beta}_t(r),$$

which corresponds to (38) for $\beta_{i,r} = \mathbf{1}\{k_i = r\}$ and $\mu_{r,t} = \underline{\beta}_t(r)$. This case is a rather special one, however, since $\beta_{i,r}$ are either 0 or 1. The grouped-factor structure thus reduces the number of parameters relative to a general factor model. This can help with all four issues of fixed-effects, since the model allows for time variation while grouping units to reduce noise. In panel event studies, grouped-factor models allow researchers to relax parallel trends and estimate heterogeneous effects of a treatment (Shin, 2022).

Computation of grouped-factor estimators can be performed using Lloyd's algorithm, suitably modified in order to allow for the time-varying structure in (42). The asymptotic properties of the method as n, T tend to infinity are similar to the time-invariant case that we reviewed in Section 5 (see Bonhomme and Manresa, 2015). In particular, under analogous conditions to the ones we discussed before, group consistency will hold, see (23).

7.4 Illustration

To explore heterogeneity in temperature effects over time, we split the observation period into three equal-sized subperiods: 1950–1968, 1969–1987, and 1988–2005. Denoting subperiods as

$p(t) \in \{1, 2, 3\}$, we then estimate a model with subperiod-specific fixed-effects, based on

$$Y_{it} = \beta_{ip(t)} X_{it} + \alpha_{ip(t)} + \delta_t + W_{it}'\gamma + \varepsilon_{it}. \tag{43}$$

In the top panel of Figure 8 we show a kernel estimator of the density of the fixed-effects estimates $\widehat{\beta}_{ip}$, for all three subperiods. In the bottom panel we plot the estimates on a map of the US, again separately by subperiod. In Appendix Table B7, columns (4) to (6), we report estimates of means, variances, and percentiles.

The fixed-effects estimates $\widehat{\beta}_{ip}$ suggest that temperature impacts have become more negative on average over time. The mean effects is -5.4 in the first subperiod, and increases to -9.6 and -8.5 in the subsequent subperiods. We observe a noticeable increase in dispersion, with standard deviations equal to 5.2, 7.9, and 9.4 in the three subperiods. We also notice that the dispersion within each subperiod is larger than that of the fixed-effects estimated on the full period.

However, as we pointed out when discussing Figure 4, a concern is that the dispersion in fixed-effects estimates partly reflects the impact of noise, which is magnified here given that the fixed effects are estimated based on shorter subpanels.

We next report estimates based on the grouped-factor model

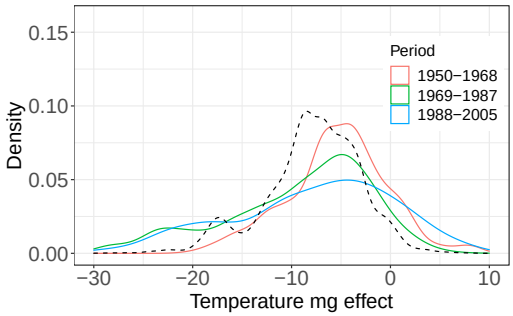
$$Y_{it} = \beta_{k(i)p(t)} X_{it} + \alpha_{ip(t)} + \delta_t + W_{it}'\gamma + \varepsilon_{it}. \tag{44}$$

In (44) we assume that group membership k_i is constant over the full period, and specify $\underline{\beta}_t(k) = \beta_{kp(t)}$ as piecewise-constant on the three subperiods. This choice is motivated by parsimony, since it limits the number of time effects to be estimated compared to a specification allowing for unrestricted group-specific time effects. In Appendix Table B8 we report estimates of means, variances, and percentiles. We show grouped-factor estimates based on $K = 4$ groups, as well as robustness checks for other numbers of groups.

We find that average effects by subperiod are close to the estimates based on fixed-effects, although they are not exactly the same. Average (negative) impacts increase from -3.9 in the first subperiod to -8.0 and -10.2 in the subsequent subperiods. The estimates also show a large increase in dispersion over time. However, the level of the dispersion is much reduced compared to the fixed-effects estimates. Dispersion is low in the first subperiod, with a standard deviation of 0.7, and it increases substantially to 3.6 and 5.2 in the subsequent ones. Figure 9 shows histograms of heterogeneous effects, and maps indicating their spatial distribution over time. Overall, estimates from the grouped-factor model (44) suggest increasingly negative temperature impacts, which become more dispersed across counties over time.

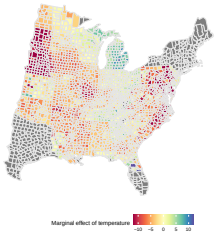
Figure 8: Fixed-effects estimates by subperiod

(a) Marginal densities

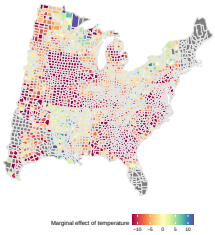


Spatial distribution

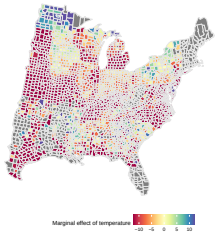
(b) 1950–1968



(c) 1969–1987



(d) 1988–2005



Notes: panel (a) shows the density of marginal effects of temperature across counties and years by period, weighted by corn area (in solid), whereas the dotted line shows the overall density based on time-invariant fixed-effects estimation. Panels (b) to (d) show the spatial distribution of marginal effects by period.

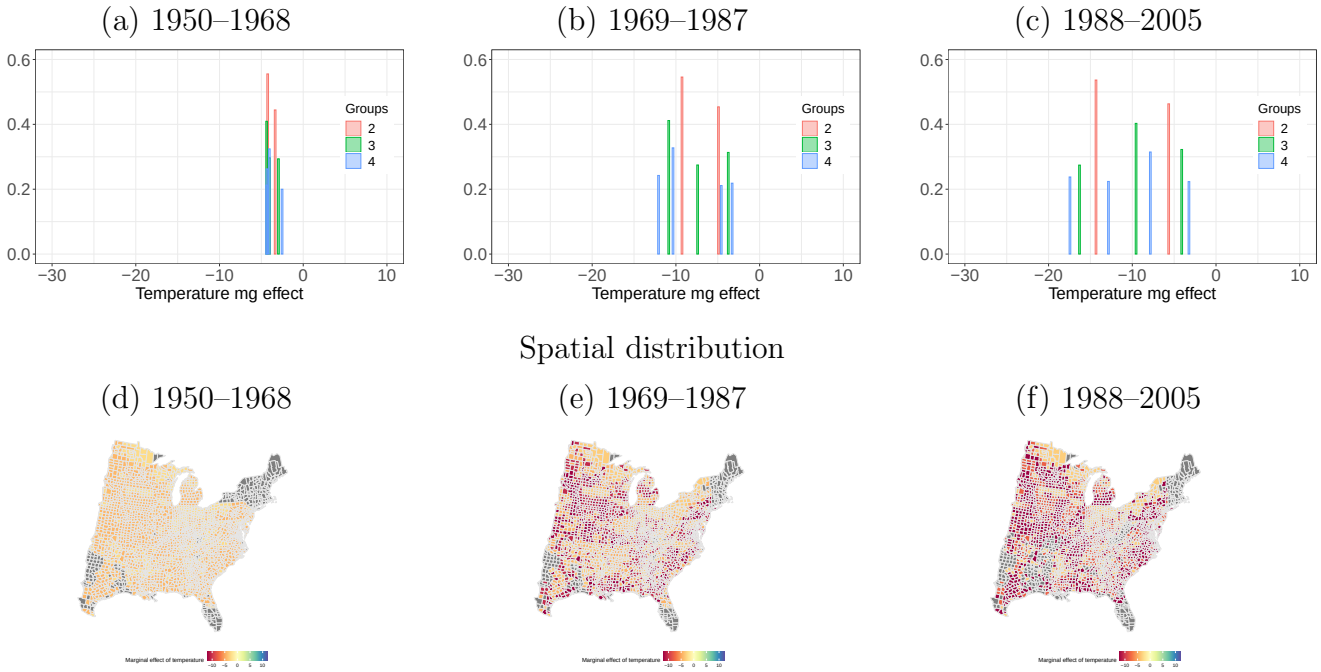
Out of sample prediction. Recall we had randomly removed one-third of the counties per state for the years 2001–2005. In the last part of this section we comment on findings from a prediction exercise that uses these held-out observations as a test sample. To proceed, we use several of the models we have estimated so far, none of which uses data from the hold-out sample, and compute the relative mean squared error

$$\sum_{(i,t)\in \text{ test}} \frac{(Y_{it} - \widehat{Y}_{it}^{\text{m}})^2}{Y_{it}^2},$$

where $\widehat{Y}_{it}^{\text{m}}$ denote the predicted outcome values under model m.

In Appendix Table B9 we report findings based on models with time-invariant heterogeneity. On average, grouped fixed-effects seems not to improve relative to fixed-effects in terms of mean squared error. However, the random-effects estimates offer some prediction gains. Next, in Appendix Table B10 we report findings based on models with time-varying heterogeneity. Allowing the fixed-effects to be subperiod-specific improves prediction substantially relative to

Figure 9: Grouped-factor estimates by subperiod



Notes: panels (a) to (c) show the histograms of marginal effects of temperature across counties and years by period, weighted by corn area, estimated by grouped-factor. Panels (d) to (f) show the spatial distribution of marginal effects by period for $K = 4$.

time-invariant fixed-effects. Moreover, allowing for temperature effects to depend on additional state-year indicators improves prediction further. While these findings should be corroborated with more systematic out-of-sample validation exercises, they suggest that accounting correctly for variation over time is important for prediction in this setting.

8 Conclusion

Better data and methods now provide applied researchers with opportunities to account for rich heterogeneity in levels and responses. A natural approach is fixed-effects estimation of heterogeneous coefficients. However, fixed-effects estimates are often too noisy. While bias-correction methods offer improvements, it is often useful to impose additional assumptions through some form of regularization. Groups, random-effects, and factor and grouped-factor methods all impose some regularization relative to models with unrestricted time-invariant or time-varying parameters. In our application, these methods shed light on the heterogeneity in

temperature impacts across space and over time.

While we have focused on linear panel data models with coefficient heterogeneity as an extension of two-way fixed-effects methods that are popular in applied work, the methods reviewed here can be used in other settings. The literature on nonlinear panel data models is now extensive, and methods of the type reviewed here can be applied to discrete choice models and other nonlinear models (e.g., Arellano and Hahn, 2007, Arellano and Bonhomme, 2011). Linear regressions on network data are now increasingly used in applications, as in the model with worker and firm fixed-effects proposed by Abowd, Kramarz, and Margolis (1999). The structure is related to the panel data models we have focused on, and methods to improve over fixed-effects are now available (e.g., Kline, Saggio, and Sølvsten, 2020, Bonhomme, Holzheu, Lamadon, Manresa, Mogstad, and Setzler, 2023).

References

Abadie, A., A. Diamond, and J. Hainmueller (2010): "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, 105(490), 493–505.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67(2), 251–333.

Andrews, D. W. (2005): "Cross-section regression with common shocks," *Econometrica*, 73(5), 1551–1585.

Arellano, M. (2003a): "Discrete choices with panel data," *Investigaciones económicas*, 27(3), 423–458.

——— (2003b): *Panel data econometrics*. OUP Oxford.

Arellano, M., and S. Bonhomme (2009): "Robust priors in nonlinear panel data models," *Econometrica*, 77(2), 489–536.

——— (2011): "Nonlinear panel data analysis," *Annu. Rev. Econ.*, 3(1), 395–424.

——— (2012): "Identifying distributional characteristics in random coefficients panel data models," *The Review of Economic Studies*, 79(3), 987–1020.

Arellano, M., and J. Hahn (2007): "Understanding bias in nonlinear panel models: Some recent developments," in *Invited Lecture, Econometric Society World Congress, London*. Citeseer.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021): "Synthetic difference-in-differences," *American Economic Review*, 111(12), 4088–4118.

Arkhangelsky, D., and G. Imbens (2024): "Causal models for longitudinal and panel data: A survey," *The Econometrics Journal*, p. utae014.

Arkhangelsky, D., and G. W. Imbens (2023): "Fixed Effects and the Generalized Mundlak Estimator," *Review of Economic Studies*, p. rdad089.

Armstrong, T. B., M. Weidner, and A. Zeleneev (2022): "Robust estimation and inference in panels with interactive fixed effects," *arXiv preprint arXiv:2210.06639*.

Arthur, D., and S. Vassilvitskii (2007): "k-means++: The advantages of careful seeding," in *Soda*, vol. 7, pp. 1027–1035.

Bai, J. (2009): "Panel data models with interactive fixed effects," *Econometrica*, 77(4), 1229–1279.

Bonhomme, S., and A. Denis (2024): "Estimating heterogeneous effects: applications to labor economics," *arXiv preprint arXiv:2404.01495*.

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (2023): "How much should we trust estimates of firm effects and worker sorting?," *Journal of Labor Economics*, 41(2), 291–322.

Bonhomme, S., T. Lamadon, and E. Manresa (2022): "Discretizing unobserved heterogeneity," *Econometrica*, 90(2), 625–643.

Bonhomme, S., and E. Manresa (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83(3), 1147–1184.

Bonhomme, S., and J.-M. Robin (2010): "Generalized non-parametric deconvolution with an application to earnings dynamics," *The Review of Economic Studies*, 77(2), 491–533.

Bonhomme, S., and M. Weidner (2022): "Posterior average effects," *Journal of Business & Economic Statistics*, 40(4), 1849–1862.

Burke, M., and K. Emerick (2016): "Adaptation to climate change: Evidence from US agriculture," *American Economic Journal: Economic Policy*, 8(3), 106–140.

Cai, J.-F., E. J. Candès, and Z. Shen (2010): "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, 20(4), 1956–1982.

Callaway, B., and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of econometrics*, 225(2), 200–230.

Chamberlain, G. (1992): "Efficiency bounds for semiparametric regression," *Econometrica: Journal of the Econometric Society*, pp. 567–596.

Chen, J. (2023): "Empirical Bayes When Estimation Precision Predicts Parameters," .

Cheng, X., F. Schorfheide, and P. Shao (2023): *Clustering for Multi-Dimensional Heterogeneity with an Application to Production Function Estimation*. Penn Institute for Economic Research, Department of Economics, University of ....

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and probability curves without crossing," *Econometrica*, 78(3), 1093–1125.

Chernozhukov, V., C. B. Hansen, Y. Liao, and Y. Zhu (2019): "Inference for heterogeneous effects using low-rank estimations," Discussion paper, CEMMAP working paper.

Chetverikov, D., and E. Manresa (2022): "Spectral and post-spectral estimators for grouped panel data models," *arXiv preprint arXiv:2212.13324*.

De Chaisemartin, C., and X. d'Haultfoeuille (2020): "Two-way fixed effects estimators with heterogeneous treatment effects," *American economic review*, 110(9), 2964–2996.

——— (2023): "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," *The Econometrics Journal*, 26(3), C1–C30.

Dell, M., B. F. Jones, and B. A. Olken (2014): "What do we learn from the weather? The new climate-economy literature," *Journal of Economic literature*, 52(3), 740–798.

Deschênes, O., and M. Greenstone (2007): "The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather," *American economic review*, 97(1), 354–385.

Dhaene, G., and K. Jochmans (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82(3), 991–1030.

Dzemski, A., and R. Okui (2024): "Confidence set for group membership," *Quantitative Economics*, 15(2), 245–277.

Efron, B. (2012): *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1. Cambridge University Press.

——— (2016): "Empirical Bayes deconvolution estimates," *Biometrika*, 103(1), 1–20.

Fernández-Val, I., and J. Lee (2013): "Panel data models with nonadditive unobserved heterogeneity: Estimation and inference," *Quantitative Economics*, 4(3), 453–481.

Fernández-Val, I., and M. Weidner (2016): "Individual and time effects in nonlinear panel models with large N, T," *Journal of Econometrics*, 192(1), 291–312.

——— (2018): "Fixed effects estimation of large-T panel data models," *Annual Review of Economics*, 10(1), 109–138.

Gobillon, L., and T. Magnac (2016): "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, 98(3), 535–551.

Goodman-Bacon, A. (2021): "Difference-in-differences with variation in treatment timing," *Journal of econometrics*, 225(2), 254–277.

Graham, B. S., and J. L. Powell (2012): "Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models," *Econometrica*, 80(5), 2105–2152.

Gu, J., and R. Koenker (2017): "Unobserved heterogeneity in income dynamics: An empirical Bayes perspective," *Journal of Business & Economic Statistics*, 35(1), 1–16.

Hahn, J., D. W. Hughes, G. Kuersteiner, and W. K. Newey (2022): "Efficient Bias Correction for Cross-section and Panel Data," *arXiv preprint arXiv:2207.09943*.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large," *Econometrica*, 70(4), 1639–1657.

Hahn, J., G. Kuersteiner, and M. H. Cho (2004): "Asymptotic distribution of misspecified random effects estimator for a dynamic panel model with fixed effects when both n and T are large," *Economics Letters*, 84(1), 117–125.

Hahn, J., and H. R. Moon (2010): "Panel data models with finite number of multiple equilibria," *Econometric Theory*, 26(3), 863–881.

Hahn, J., and W. Newey (2004): "Jackknife and analytical bias reduction for nonlinear panel models," *Econometrica*, 72(4), 1295–1319.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): "Statistical learning with sparsity," *Monographs on statistics and applied probability*, 143(143), 8.

Hsiao, C., and M. H. Pesaran (2008): "Random coefficient models," in *The econometrics of panel data: Fundamentals and recent developments in theory and practice*, pp. 185–213. Springer.

Jochmans, K., and M. Weidner (2024): "Inference on a distribution from noisy draws," *Econometric Theory*, 40(1), 60–97.

Keane, M., and T. Neal (2020): "Climate change and US agriculture: Accounting for multidimensional slope heterogeneity in panel data," *Quantitative Economics*, 11(4), 1391–1429.

Kiefer, J., and J. Wolfowitz (1956): "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, pp. 887–906.

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-out estimation of variance components," *Econometrica*, 88(5), 1859–1898.

Koenker, R., and J. Gu (in preparation): *Empirical Bayes: Some Tools, Rules and Duals*. Econometric Society Monographs.

Koenker, R., and I. Mizera (2014): "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, 109(506), 674–685.

Kuersteiner, G. M., and I. R. Prucha (2013): "Limit theory for panel data models with cross sectional dependence and sequential exogeneity," *Journal of Econometrics*, 174(2), 107–126.

——— (2020): "Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity," *Econometrica*, 88(5), 2109–2146.

Li, T., and Q. Vuong (1998): "Nonparametric estimation of the measurement error model using multiple indicators," *Journal of Multivariate Analysis*, 65(2), 139–165.

Lin, C.-C., and S. Ng (2012): "Estimation of panel data models with parameter heterogeneity when group membership is unknown," *Journal of Econometric Methods*, 1(1), 42–55.

Lu, X., and L. Su (2017): "Determining the number of groups in latent panel structures with an application to income and democracy," *Quantitative Economics*, 8(3), 729–760.

Miller, S., K. Chua, J. Coggins, and H. Mohtadi (2021): "Heat waves, climate change, and economic output," *Journal of the European Economic Association*, 19(5), 2658–2694.

Moon, H. R., and M. Weidner (2018): "Nuclear norm regularized estimation of panel regression models," *arXiv preprint arXiv:1810.10987*.

Mugnier, M. (2022): "A simple and computationally trivial estimator for grouped fixed effects models," Discussion paper, Working paper.

Pesaran, M. H. (2006): "Estimation and inference in large heterogeneous panels with a multifactor error structure," *Econometrica*, 74(4), 967–1012.

Pesaran, M. H., and R. Smith (1995): "Estimating long-run relationships from dynamic heterogeneous panels," *Journal of econometrics*, 68(1), 79–113.

Pollard, D. (1981): "Strong consistency of k-means clustering," *The annals of statistics*, pp. 135–140.

——— (1982): "A central limit theorem for k-means clustering," *The Annals of Probability*, 10(4), 919–926.

Rambachan, A., and J. Roth (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90(5), 2555–2591.

Schlenker, W., and M. J. Roberts (2009): "Nonlinear temperature effects indicate severe damages to US crop yields under climate change," *Proceedings of the National Academy of sciences*, 106(37), 15594–15598.

Shin, M. (2022): "Finitely Heterogeneous Treatment Effect in Event-study," *arXiv preprint arXiv:2204.02346*.

Stefanski, L. A., and R. J. Carroll (1990): "Deconvolving kernel density estimators," *Statistics*, 21(2), 169–184.

Su, L., Z. Shi, and P. C. Phillips (2016): "Identifying latent structures in panel data," *Econometrica*, 84(6), 2215–2264.

Sun, L., and S. Abraham (2021): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of econometrics*, 225(2), 175–199.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.

Wooldridge, J. M. (2005): "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models," *Review of economics and statistics*, 87(2), 385–390.

Yu, L., J. Gu, and S. Volgushev (2022): "Group structure estimation for panel data–a general approach," *arXiv preprint arXiv:2201.01793*.

——— (2024): "Spectral clustering with variance information for group structure estimation in panel data," *Journal of Econometrics*, 241(1), 105709.

ONLINE APPENDIX

A Appendix: data

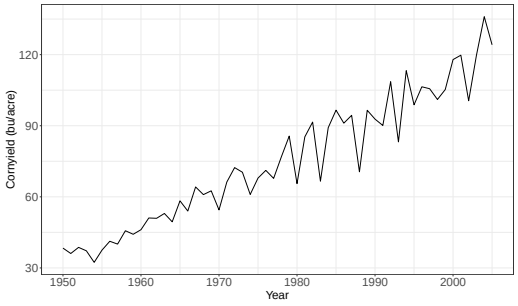
Table A1: Descriptive statistics

	Obs	P10	P25	Median	Mean	P75	P90	St. Dev.
Temperature (C)	104,149	16.2	17.9	20.2	20.2	22.4	24.2	3.0
Precipitation (mm)	104,149	2.3	2.8	3.3	3.4	3.9	4.5	0.8
Corn yields (bu/acre)	104,149	31.2	46.0	71.0	74.2	97.7	122.6	34.7
Corn area (1,000s acres)	104,149	0.8	2.6	11.9	32.3	45.4	96.8	44.9

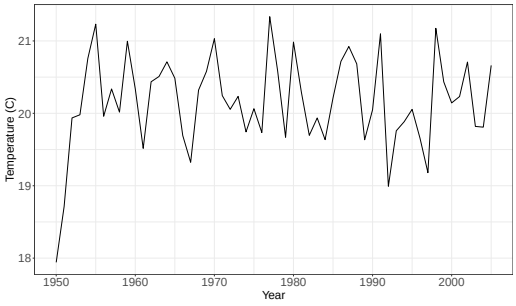
Notes: un-weighted statistics.

Figure A1: Descriptive figures of baseline sample

(a) Mean yields



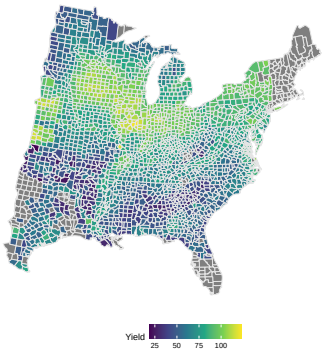
(b) Mean temperature



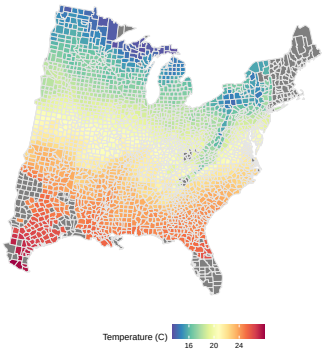
Notes: un-weighted statistics across counties. Yields are measured in bushels per acre. Temperature is measured in average daily degree Celsius above zero degrees Celsius during the growing season.

Figure A2: Maps of averages per county

(a) Yields



(b) Temperature



Notes: un-weighted means (1950–2005).

B Appendix: robustness

B.1 Alternative standard errors

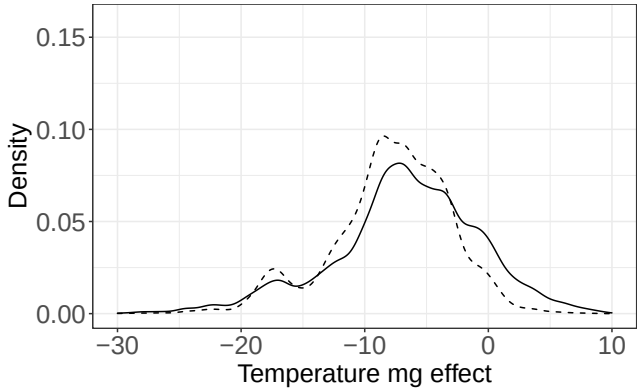
Table B2: Yield regressions - Driscoll and Kraay standard errors

	2 Lags				3 Lags			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Temperature	-3.924	-3.488	-3.875	-6.600	-3.924	-3.488	-3.875	-6.600
	(0.379)	(3.968)	(1.418)	(0.926)	(0.372)	(3.882)	(1.404)	(0.942)
Precipitation	5.664	7.244	3.097	1.991	5.664	7.244	3.097	1.991
	(1.879)	(2.162)	(0.743)	(0.382)	(1.972)	(2.276)	(0.741)	(0.398)
Observations	104,149	104,149	104,149	104,149	104,149	104,149	104,149	104,149
County FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes	No	No	Yes	Yes
State-year FE	No	No	No	Yes	No	No	No	Yes

Notes: un-weighted regressions. Driscoll and Kraay standard errors with small sample correction.

B.2 Results without weights

Figure B3: Fixed-effects results - Density of marginal effects un-weighted

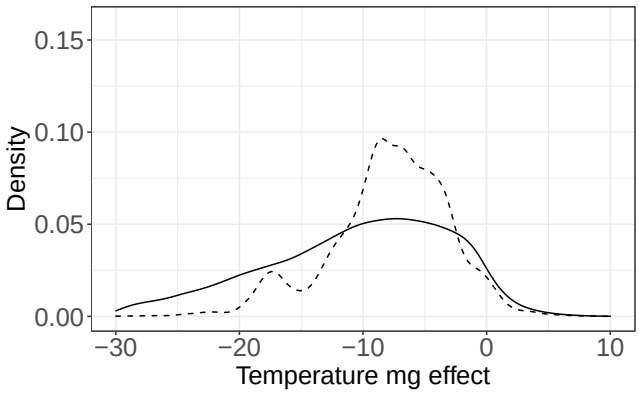


Notes: density of marginal effects across counties and years estimated via fixed-effects. The solid line shows the un-weighted results, while the dashed line presents the results weighted by corn area.

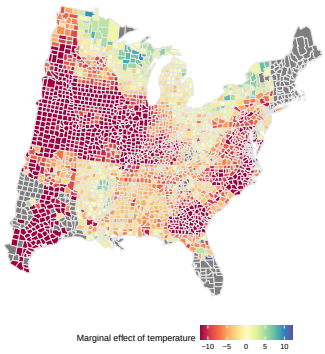
B.3 Regression of log yields

Figure B4: Fixed-effects results in a regression of log yields

(a) Marginal density



(b) Spatial distribution



Notes: panel (a) shows the density of marginal effects estimated via fixed-effects in a regression for yields in levels (dotted line) versus in logs (solid line). Panel (b) plots marginal effects per county from a regression in logs. For the regression in logs, errors are assumed to be normally distributed and marginal effects are estimated as $\widehat{\beta}_i \cdot exp(\widehat{log(y_{it})}) \cdot exp(\widehat{\sigma}_\varepsilon^2/2)$.

B.4 Growing and killing degree days

The literature has studied nonlinear effects of temperature by distinguishing between *growing degree days* (gdd), measured as degree-days between 0 and 29C, and *killing degree days* (kdd), measured as degree-days above 29C, both measured during the growing season (see Keane and Neal, 2020).[1] As with our main temperature variable, we re-scale these variables by the number of days in the growing season.

We estimate the following specification,

$$y_{it} = \alpha_i + \beta_i gdd_{it} + \gamma_i kdd_{it} + \chi precr_{it} + \lambda_{s(i)t} + \epsilon_{it}, \tag{B1}$$

where $precr_{it}$ denote precipitations, and $\lambda_{s(i)t}$ include state-year fixed-effects. Note that equation (B1) does not correspond to a piece-wise linear specification based on our baseline measure of temperature, because the daily distribution of temperature matters as well. That is, two counties may have the same overall degree days, but with different values for *growing* and *killing degree days*.

Table B3 shows descriptive statistics for our baseline temperature measure (DD_{0+}), and the measures of growing degree days ($GDD_{0,29}$) and killing degree days (KDD_{29+}).

Using (B1), along with our baseline fixed-effects specification, we estimate the expected change in yields of a daily increase in temperature of 1 degree Celsius. Table B4 shows the distribution of marginal effects associated with an increase in 1 degree Celsius daily. In column (1) we report our main estimates based on a model with time-invariant β_i. In column (2) we report estimates implied by the nonlinear model (B1). We see that the nonlinear model implies a smaller average effect, and a somewhat higher dispersion of effects. In Figure B5 we plot the density of effects. In Figure B6 we plot the effects on a map of the US. We see that the spatial distribution is quite similar in the two models.

However, the effects of an increase in growing degree days or killing degree days are very different, and differently distributed across space. To illustrate this, in Figure B7 we plot the county-specific coefficients of growing and killing degree days in (B1), respectively.

[1] Keane and Neal (2020) estimate a specification in logs and consider the growing season to be between May 1st to September 30th, while we focus on a specification in levels and, following Deschênes and Greenstone (2007) and Burke and Emerick (2016), consider the growing season to be between April 1st to September 30th.

Table B3: Various measures of temperature

	Baseline temperature			Counterfactual temperature		
	DD_{0+}	$GDD_{0,29}$	KDD_{29+}	DD_{0+}	$GDD_{0,29}$	KDD_{29+}
Percentile 10	16.460	16.396	0.039	17.460	17.325	0.079
Percentile 25	17.421	17.319	0.079	18.421	18.234	0.141
Percentile 50	18.614	18.444	0.154	19.614	19.341	0.247
Percentile 75	20.020	19.754	0.289	21.020	20.620	0.426
Percentile 90	21.840	21.413	0.486	22.840	22.227	0.678
Mean	18.922	18.703	0.219	19.922	19.585	0.326
Variance	4.771	4.054	0.044	4.771	3.846	0.075

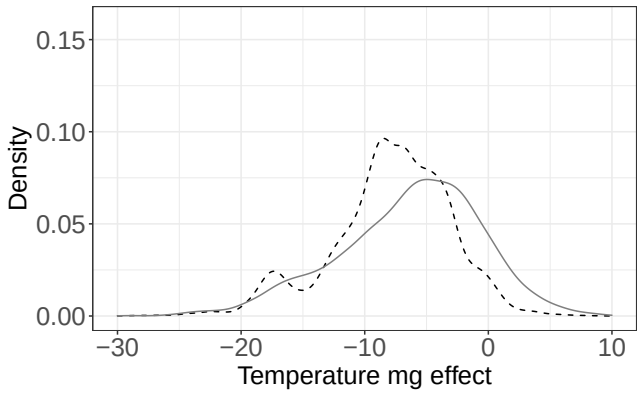
Notes: distribution across counties and years, weighted by corn area. Variables are re-scaled by 1/183, where 183 are the number of days in the growing season. DD_{0+} corresponds to our baseline temperature measure used in the main text.

Table B4: Distribution of effects of increasing daily temperature by 1C

	Baseline model	Nonlinear model
	(1)	(2)
Percentile 10	-14.701	-14.611
Percentile 25	-10.255	-9.897
Percentile 50	-7.416	-5.674
Percentile 75	-4.574	-2.265
Percentile 90	-2.356	0.521
Mean	-7.856	-6.307
Variance	24.850	36.109

Notes: density of effects of a daily increase in temperature by 1C across counties and years, weighted by corn area. For our baseline model in column (1), the effects correspond to the coefficients β_i. For the nonlinear specification in column (2), see (B1), the effects correspond to the difference of the counterfactual and baseline temperatures, gdd_{it} and kdd_{it}, weighted by the corresponding coefficients.

Figure B5: Density of effects of increasing daily temperature by 1C



Notes: density across counties and years, weighted by corn area. The solid line shows the nonlinear results from specification (B1), while the dashed line presents the baseline results.

Figure B6: Maps of marginal effects of increasing daily temperature by 1C

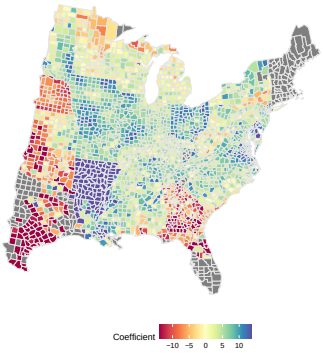
(a) Baseline model

(b) Nonlinear model

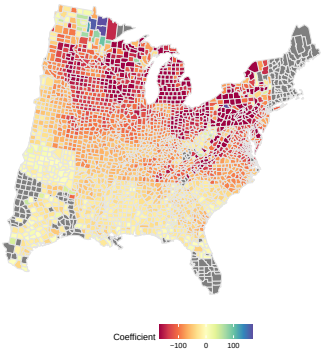
Notes: for our baseline model in panel (a), the effects correspond to the coefficients β_i. For the nonlinear specification in panel (b), see (B1), the effects correspond to the difference of the counterfactual and baseline temperatures, gdd_{it} and kdd_{it}, weighted by the corresponding coefficients.

Figure B7: Maps of coefficients in the nonlinear specification

(a) Growing degree days



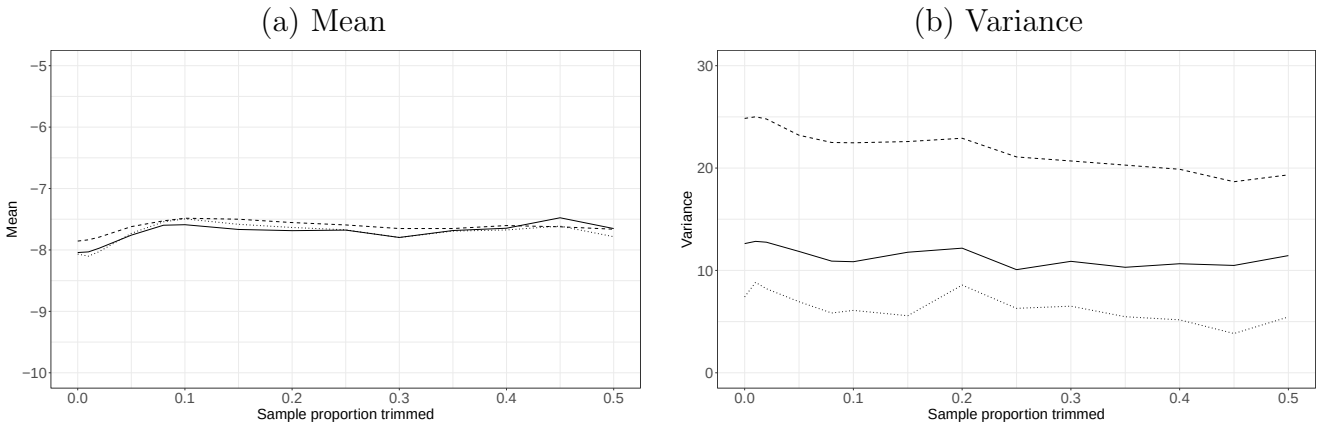
(b) Killing degree days



Notes: coefficients of gdd_{it} (left panel) and kdd_{it} (right panel) in the nonlinear specification, see (B1).

B.5 Sensitivity to trimming in Jackknife

Figure B8: Sensitivity to trimming of Jackknife results



Notes: dashed lines show results from fixed-effects estimates, solid lines show results from jackknife estimates of the distribution function, and the dotted lines show the results from the jackknife formula applied to the mean and variance directly.

B.6 Grouped fixed-effects estimates

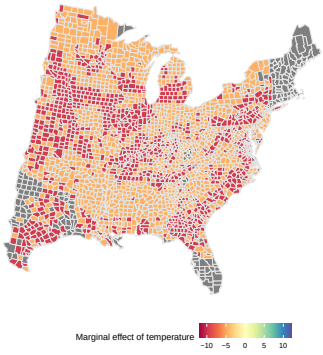
Table B5: Grouped fixed-effects, distribution of marginal effects

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$
Percentile 10	-9.690	-11.546	-14.648	-16.004	-12.464	-11.208	-12.184
Percentile 25	-9.690	-11.546	-9.141	-10.731	-9.270	-8.222	-9.170
Percentile 50	-9.690	-7.089	-9.141	-7.409	-6.731	-8.222	-6.813
Percentile 75	-4.621	-7.089	-5.549	-4.175	-3.910	-5.600	-4.261
Percentile 90	-4.621	-3.048	-1.732	-4.175	-3.910	-2.722	-1.448
Mean	-7.424	-7.573	-7.604	-7.796	-7.936	-7.560	-7.817
Variance	6.352	9.179	12.481	16.085	19.152	15.996	18.831

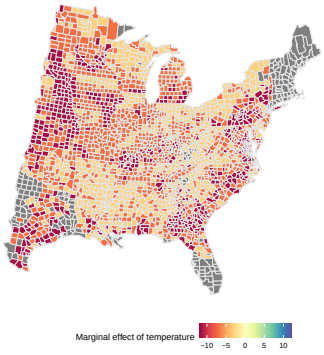
Notes: distribution of marginal effects across counties and years, weighted by corn area.

Figure B9: Grouped fixed-effects robustness, maps of marginal effects per county

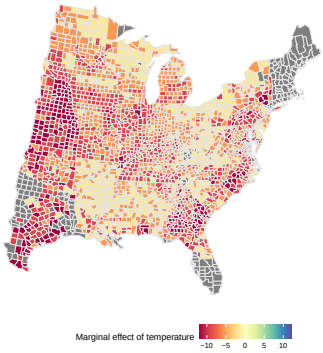
$K = 2$



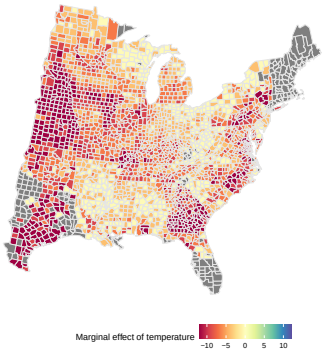
$K = 3$



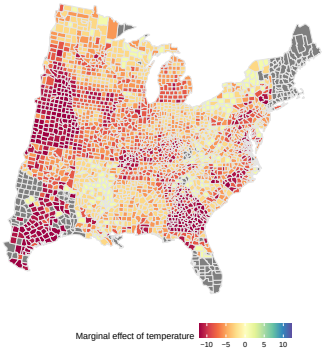
$K = 4$



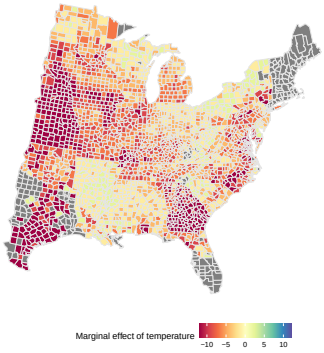
$K = 6$



$K = 7$



$K = 8$



B.7 Random-effects estimates

Table B6: Random-effects estimates – robustness

	Uncorrelated RE	Correlated RE			Posterior means
	(1)	(2)	(3)	(4)	(5)
Percentile 10	-13.678	-13.745	-13.678	-13.665	-13.482
Percentile 25	-10.425	-10.452	-10.292	-10.358	-9.916
Percentile 50	-7.612	-7.625	-7.519	-7.585	-7.607
Percentile 75	-5.012	-5.025	-5.079	-5.025	-5.207
Percentile 90	-2.786	-2.826	-3.039	-2.826	-3.658
Mean	-7.904	-7.950	-7.962	-7.924	-7.950
Variance	18.026	18.221	18.563	18.922	14.394

Notes: distribution of marginal effects across counties and years, weighted by corn area. All results are based on a normal prior. Column (1) presents the uncorrelated random-effects results. Columns (2) to (4) present correlated random-effects results under different assumptions. In column (2), the prior mean is correlated with temperature. In columns (3) and (4) the prior variance is also correlated with temperature; in column (3) via an exponential specification and in column (4) via a discrete specification with 3 bins. Column (5) presents the distribution of the posterior means corresponding to the prior assumed in column (2).

B.8 Variation over time

Table B7: Distribution of marginal effects – fixed-effects-per-period estimates

	Baseline fixed effects			Fixed effects per period		
	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3
	(1)	(2)	(3)	(4)	(5)	(6)
Percentile 10	-14.555	-14.506	-15.250	-12.444	-21.930	-21.652
Percentile 25	-10.214	-10.192	-10.468	-8.161	-13.810	-13.799
Percentile 50	-7.447	-7.402	-7.424	-5.091	-7.777	-6.891
Percentile 75	-4.575	-4.598	-4.494	-2.174	-3.967	-1.875
Percentile 90	-2.302	-2.532	-2.327	0.902	-1.273	2.086
Mean	-7.845	-7.843	-7.879	-5.370	-9.599	-8.457
Variance	24.008	24.740	25.849	26.567	61.926	88.175

Notes: statistics are weighted using corn area. In this table, we drop 113 observations corresponding to counties with one observation in one of the periods.

Table B8: Distribution of marginal effects - grouped-factor estimates

	$K = 2$			$K = 3$			$K = 4$		
	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Percentile 10	-4.246	-9.273	-14.335	-4.370	-10.856	-16.306	-4.371	-12.074	-17.447
Percentile 25	-4.246	-9.273	-14.335	-4.370	-10.856	-16.306	-4.345	-10.348	-12.848
Percentile 50	-4.246	-9.273	-14.335	-4.027	-7.406	-9.549	-4.053	-10.348	-7.847
Percentile 75	-3.354	-4.917	-5.654	-2.961	-3.747	-4.097	-4.053	-4.585	-7.847
Percentile 90	-3.354	-4.917	-5.654	-2.961	-3.747	-4.097	-2.528	-3.284	-3.235
Mean	-3.849	-7.295	-10.313	-3.854	-7.680	-9.645	-3.892	-8.004	-10.220
Variance	0.197	4.704	18.742	0.352	9.024	22.107	0.484	13.157	26.637

Notes: statistics are weighted using corn area.

B.9 Prediction

We compare predictions in the hold-out sample for each of the estimated models.[2]

Table B9: Distributions of relative squared errors for models with time-invariant slopes

	FE	GFE							RE
		$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	post means
Percentile 10	0.015	0.013	0.013	0.012	0.014	0.014	0.012	0.015	0.013
Percentile 25	0.100	0.102	0.100	0.107	0.102	0.105	0.103	0.103	0.095
Percentile 50	0.436	0.446	0.440	0.445	0.448	0.431	0.432	0.441	0.397
Percentile 75	1.388	1.429	1.419	1.435	1.415	1.405	1.416	1.421	1.262
Percentile 90	3.911	3.816	3.832	3.872	3.872	3.853	3.847	3.881	3.463
Mean	2.608	2.618	2.613	2.603	2.608	2.598	2.600	2.596	2.373
Variance	292.830	295.142	291.724	287.264	286.659	284.771	285.816	284.081	259.315

Notes: relative squared errors are defined as $(Y_{it} - \widehat{Y}_{it}^m)^2/Y_{it}^2 \times 100$*, where* $\widehat{Y}_{it}^m$ *corresponds to predicted yields under model* m*. Distribution in hold-out sample, at the county-year level, weighted by corn area. FE is fixed-effects, GFE is grouped fixed-effects, RE is random-effects. For the prediction under random-effects, we estimate all coefficients except for* β_i *by OLS in a regression where the left-hand side is* $Y_{it} - \widehat{\beta}_i^{\text{PM}} X_{it}$*.*

[2]The hold-out sample has 2,855 observations but we exclude 10 observations in two counties that do not have a predicted value for at least one of the models.

Table B10: Distributions of relative squared errors for models with time-varying slopes

	FE μ_t	GFE μ_t			FE pp	GFE pp			FE st-yr	FE st-yr
		$K=4$	$K=5$	$K=6$		$K=2$	$K=3$	$K=4$	α_i	$\alpha_{ip(t)}$
Percentile 10	0.018	0.019	0.018	0.018	0.019	0.020	0.016	0.018	0.012	0.009
Percentile 25	0.097	0.101	0.094	0.094	0.094	0.130	0.114	0.119	0.066	0.075
Percentile 50	0.445	0.419	0.424	0.424	0.402	0.457	0.441	0.438	0.328	0.318
Percentile 75	1.363	1.375	1.373	1.373	1.323	1.355	1.441	1.465	1.040	1.053
Percentile 90	3.867	3.879	3.880	3.880	3.642	3.464	3.573	3.654	3.057	3.027
Mean	2.571	2.554	2.543	2.543	2.170	2.164	2.122	2.135	1.921	1.684
Variance	283.673	270.780	270.212	270.212	160.351	147.163	131.540	136.740	134.895	86.178

Notes: relative squared errors are defined as $(Y_{it} - \widehat{Y}_{it}^m)^2/Y_{it}^2 \times 100$, *where* $\widehat{Y}_{it}^m$ *corresponds to predicted yields under model* m. *Distribution in hold-out sample, at the county-year level, weighted by corn area. FE* μ_t *is fixed-effects based on (37). GFE* μ_t *is a similar specification, where* β_i *in (37) is discrete with* K *groups. FE pp is fixed-effects per period (43). GFE pp is grouped-factor with* K *groups, see (44). FE st-yr is fixed-effects in an augmented version of (37) with state-year indicators. In the next-to-last column we include constant county-specific intercepts, while in the last column we include period-and-county-specific intercepts.*

C Implementation

In this appendix we describe how we implemented the various estimation methods.

C.1 Fixed-effects

We run fixed-effects regressions in R using the package *fixest*. For the short-panel results displayed in Figure 4, we further restrict the sample to years from 1990 to 2005 and counties with at least 10 observations in those years.

C.2 Jackknife

We initially applied the half-panel jackknife approach of Dhaene and Jochmans (2015) to estimate the mean, variance, and distribution function of β_i. However, the estimated distributions of β_i in these two samples are quite different, suggesting miss-specification, possibly due to the lack of stationarity given the presence of state-year fixed-effects. This motivated us to apply instead the bias-correction approach proposed by Fernández-Val and Weidner (2016). For the cross-section, we randomly split counties within each state into two (weighted) halves, 5 times, and average across them. For the panel dimension, we split the panel in half.

We apply this approximate bias-correction formula to estimate the distribution of β_i, that is, to estimate $\widehat{F}_\beta(b)$ for a vector of b values.[3] However, this results in an estimated function that is not monotone and not always bounded between 0 and 1. We address the first issue by rearrangement (see Chernozhukov, Fernández-Val, and Galichon, 2010),[4] and we then truncate the values to the unit interval. With this estimated distribution function, we simulate data to generate the density shown in Figure 5.

Note that applying the bias-correction approach directly to quantiles gives different estimates than estimating the quantiles from the bias-corrected distribution. Similarly, by applying the correction to the mean and the variance we obtain different point estimates than the mean and variance implied by the bias-corrected distribution. In the case of the mean, both values are similar, -8.07 and -8.03 respectively, while in the case of the variance the discrepancy is larger, with values of 7.4 and 12.5 respectively. This pattern also holds when we trim counties with the lowest variability in temperature, see Figure B8.

[3]The fixed-effects estimates $\widehat{\beta}_i$ range between -39.3 and 14. We define a vector b of length 10,001 equally-spaced between -45 and 45. In each subsample, we estimate the empirical distribution function at each value of b, weighted by corn area.

[4]We use the function *stepfun* from the R package *quantreg*.

C.3 Grouped fixed-effects

To estimate grouped fixed-effects, given a number of groups K, we apply Lloyd's algorithm to de-meaned outcomes and covariates (which is equivalent to estimating α_i in (18)), and estimate $k_1, ..., k_n, \delta, \gamma, \underline{\beta}$. We use multiple starting values $\delta, \gamma, \underline{\beta}$ to initialize the algorithm. As a first starting value we use the fixed-effects estimates, group $\widehat{\beta}_i$ into K groups and take the average value within each group for $\underline{\beta}$. For the subsequent starting values, we take draws from a normal distribution centered at the first starting value, and with a diagonal covariance matrix. For the variance of δ and γ, we take the squared standard error of each coefficient estimated from fixed-effects, which are clustered at the state level, multiplied by 10. For each $\underline{\beta}_k$, we take the variance of the fixed-effects estimates across counties.

In our implementation, we notice that choosing a variance for the heterogeneous parameters that is too large may worsen the outcome of the algorithm. If the $\underline{\beta}_k$'s take implausible values, in the first step of the first iteration all units are assigned to one group, and therefore the other $\underline{\beta}_k$'s are not estimated in the second step of that first iteration. This implies that in the next iteration, the residuals can only be estimated for one group, all units are assigned to that group (which is the same group as in the previous iteration), and therefore the coefficients are not updated and the algorithm stops. One way to avoid this problem is to have reasonable, not too large starting values for the $\underline{\beta}$'s.

The estimates we present are based on 3,000 starting values, 10,000 iterations for each one of them, and a tolerance level for the change in the objective function between iterations of $1e^{-7}$. We estimate grouped fixed-effects for values of K from 2 to 8. Column (4) in Table 2 presents the results for $K = 5$, which corresponds to the number of groups chosen by the information criterion proposed by Su, Shi, and Phillips (2016), while Appendix Table B5 presents the results for other values of K.[5]

C.4 Random-effects

To estimate the variance V_i of fixed-effects estimates, we use the squared standard errors of $\widehat{\beta}_i$ under several assumptions: i.i.d. homoskedastic, heteroskedastic, Newey-West with one lag, and clustered at the county level. We have found that the random-effects estimates are similar across all the specifications. We report the results assuming i.i.d. homoskedastic errors.

[5] As a check for the computation, we found it useful to compare the distribution of β_i estimates obtained from the best 10 starting values, corresponding to those that give the 10 minimum values of the objective function. We found stable estimates for all $K \leqslant 6$, but more variability for $K = 7, 8$, suggesting that these estimates may be less reliable.

We entertain several specifications for the priors on β_i. First, we assume each β_i is an independent draw from the same distribution $\mathcal{N}(\mu, \sigma^2)$, i.e. we assume independent random-effects. Second, we assume β_i are i.i.d. normal with a mean that depends linearly on the average temperature of county i. In further specifications, we allow for the variance to also depend on temperature, using two different specifications: an exponential model, and a discrete model with 3 bins. In the discrete case, we group counties in 3 groups according to average temperature.

C.5 Factors and grouped factors

We implement two versions with time-varying coefficients α_{it} and β_{it}, allowing them to be period-specific for 3 subperiods, 1950–1968, 1969–1987, and 1988–2005. In the first version, we allow the coefficients β_{it} to be unrestricted within period across counties, while in the second one we assume that counties are grouped, but their effects may vary by period. In both cases, we allow the intercepts to be fully heterogeneous across counties and within periods, $\alpha_{ip(t)}$. Note that, when allowing temperature effects to be county-specific but varying across periods, we loose 113 observations corresponding to counties-periods that have only one observation.

For the grouped-factor version, we assume groups are fixed across periods. The implementation is similar to the time-invariant case, except that we de-meaned observations at the county-period level (instead of de-meaning at the county level), and that the number of parameters to estimate increases.[6] Table B8 describes the results for $K = 2, 3, 4$.

[6] We also allow for the coefficients in γ not to be identified, by modifying the code such that the coefficients are set to zero in case of multicollinearity. This technical modification only matters in intermediate steps of Lloyd's algorithm.