


Food and agricultural industry locational determinants research: Aggregation bias and size measurement in the agricultural support industry

Craig Wesley Carpenter^{1*} , Anders Van Sandt² and Scott Loveridge¹

¹Department of Agricultural, Food, and Resource Economics, Michigan State University, East Lansing, MI, USA and ²Department of Agricultural and Applied Economics, University of Wyoming, Laramie, WY, USA
*Corresponding author. Email: cwcarp@msu.edu

(Received 5 April 2022; revised 6 September 2022; accepted 6 September 2022; first published online 11 November 2022)

Abstract

Federal administrative data present a valuable opportunity for food and agricultural industry locational outcome research. We review issues with aggregated U.S. public data and summarize current methods. An example empirical approach combines federal administrative and secondary data. We compare results with differing levels of industrial aggregation. Results indicate locational determinants vary in magnitude, sign, and significance across industries and their sub-industries, as well as between employers and non-employers – nuances commonly missed with public data. We conclude by emphasizing that studies relying on public (more-aggregated) data may miss locational outcome relationships or inappropriately generalize to sub-industries and suggest data access changes.

Keywords: administrative data; census data; FAI; farm services; locational choice

Introduction

For decades, U.S. research and policy targeted farm retention (Ahearn, Yee, and Korb 2005; Akobundu et al. 2004; Goetz and Davlasheridze 2017). However, an increasingly large proportion of the U.S. food and fiber supply chain is off-farm and encompasses economic activity such as producer services, manufacturing, transportation, warehousing, wholesaling, and retail trade, extending beyond traditional agricultural industrial classifications. The agricultural supply chain is an important component of rural economies, resulting from rural and often region-specific production of agricultural commodities. Productivity and locational changes in off-farm food and agricultural industry (FAI) sectors will influence on-farm activity, and the resulting effects, upstream and downstream from the farm, can have important feedback effects on agricultural producers. Despite their importance, the attributes associated with FAI locations along the agricultural supply

chain, such as agricultural support services and farm-related manufacturers and wholesalers, have not been thoroughly explored.

A key barrier to modeling FAI industry locational outcomes has been the availability of comprehensive and accurate data, particularly for smaller rural economies. The relatively limited county-level counts of agriculture-related businesses in rural areas lead to heavy suppression of publicly available data on these establishments, resulting in either incomplete, inaccurate, or biased data and data estimates (Carpenter, Van Sandt, and Loveridge 2022c). Since agricultural support services are often geographically bound by the type of crop they support, the heavily censored publicly available data sets that are available contain many zeros, further complicating the process of modeling their locational outcomes. Researchers have hitherto been constrained to use either highly aggregated data or estimated point data of unknown error.

The objective of this article is to demonstrate the potential value of modeling FAI establishment locations using restricted access microdata on multiple industry size measures in the Federal Statistical Research Data Center (FSRDC) system.¹ To achieve this objective, we use FSRDC restricted access data to first econometrically estimate the determinants of FAI locational outcomes for Support Activities for Agriculture and Forestry (North American Industrial Classification System [NAICS] 115), a relatively aggregated industry that is commonly used in response to public data suppressions (Hertz and Zahniser 2013).² Then, we examine its subsectors – Support Activities for Crop Production (NAICS 1151), Support Activities for Animal Production (NAICS 1152), Support Activities for Forestry (NAICS 1153) – using the same approach. The FSRDC data make several improvements over public or proprietary data, which have limitations that are especially constraining in locational outcome research.³ Much of the literature revolving around business locational outcomes measures industry size by the number of establishments within an industry and region (Carpenter, Van Sandt, and Loveridge 2021).

Several studies explored these measures in other industries at higher levels of industry aggregation, but the confidential nature of this data has restricted its use (Cleary et al. 2018). Data aggregations and suppressions not only limit the understanding of industries germane to the agricultural supply chain but also constrain agricultural support efforts and rural economic development. As shown in the results, aggregation across related sectors can mask (and even reverse the direction of) important statistical relationships. In an effort to overcome these issues, researchers often resort to proprietary data sets with unpublished suppressed cell estimation techniques, or to estimating suppressed cells with their own

¹This article follows the U.S. Census Bureau convention of using the term “establishment” to refer to a physical location or “address” where economic activity takes place and the term “firm” to refer to a collection of one or more establishments under a common ownership structure.

²The term “locational outcomes” serves as an umbrella term to encompass both demand threshold analysis and business location research. The sections below elaborate on the differences in the related theoretical perspectives, but this terminology avoids arguing the example models below have a causal interpretation, while also emphasizing the broad potential of future Federal microdata FAI research.

³NAICS 1151 industries include Cotton Ginning (NAICS 115111), Soil Preparation, Planting, and Cultivating (115112), Crop Harvesting, Primarily by Machine (115113), Postharvest Crop Activities (except Cotton Ginning) (115114), and Farm Labor Contractors and Crew Leaders (115115), and Farm Management Services (115116). NAICS 1152 includes industries such as establishments engaged in breeding services, boarding horses, livestock spraying, and sheep dipping and shearing. And NAICS 1153 includes industries establishment engaged in such as aerial forest mulching or seeding, cruising timber, estimating and valuating timber, and forest pest control services. See the North American Industrial Classification System from the United States Office of Management and Budget for additional details.

algorithms. Both solutions yield biased estimates (Carpenter, Van Sandt, and Loveridge 2022c).

The Longitudinal Business Database (LBD) is an annual series produced by the U.S. Census Bureau based on establishment records from the Business Register (Jarmin and Miranda 2002). The LBD remains a core data set for studying the characteristics and determinants of entry, growth, and exit at the establishment, firm, industry, and economy-wide level (Carpenter and Loveridge 2018, 2019, 2020; Davis, Faberman, and Haltiwanger 2006; Foster, Haltiwanger, and Krizan 2006; Haltiwanger, Jarmin, and Miranda 2013). The Integrated Longitudinal Business Database (ILBD) covers non-employers.⁴ Longitudinal establishment-level data that encompass an entire state or regional economy are similarly becoming available in other countries, with economists often using them to examine firm establishment locations, though FAI locational analysis remains relatively neglected (Arauzo-Carod and Viladecans-Marsal 2009; Chen and Moore 2010; Devereux, Griffith, and Simpson 2007; Figueiredo, Guimaraes, and Woodward 2002; Holl 2004).

An FSRDC is thus a natural place to make improvements to research. Researchers cannot release minimum or total counts, but we develop unbiased coefficients in regression results. By using the LBD, ILBD, and publicly available county-level data, we propose improving existing FAI location research by:

1. Increasing accuracy in industry size measurement by comparing location determinants of both employer establishments and non-employer establishments.
2. Reducing generalizability by expanding geographic scope.
3. Testing the influence of industry aggregation on FAI location outcome models.

The next section summarizes prior research on FAI locational outcomes. Then, we review related econometric and data issues. This review leads to our flexible approach in the supporting example empirical method, which uses FSRDC data. The example results section demonstrates that significant location attributes of FAI vary depending on the type and level of industry in the FAI supply chain. We note different data generation processes not only across different FAI but also between employers and non-employers within a specific industry.⁵ In most FAI, agricultural production variables are important contributors to the likelihood of a county containing at least some FAI establishments, as well as the number of establishments. We also show that aggregating sectors can lead to substantial changes in the magnitude, significance, and even direction of estimated coefficients. The article concludes by summarizing our review and supporting modeling exercise to achieve our objective of demonstrating the value of FSRDC data and flexible modeling in accurately modeling FAI locational outcomes.

⁴The ILBD is less refined than the LBD and is not without criticism. The limited revisions imply that many of the records resulting from tax forms and (otherwise identical) operations potentially appearing as single or multiple entities, depending on tax and liability decisions. Nonetheless, the ILBD remains a substantial improvement over publicly available datasets and we argue it remains useful, especially when aggregated. Further, the limitations of the ILBD provide further justification to keep the non-employer measurements separate from the employer measurements.

⁵Non-employer establishments are establishments that operate without paid employees, often family or single sole proprietor establishments.

FAI locational research

Using FSRDC microdata, Dunn and Hueth (2017) show the number of crop services establishments is in secular decline, but the number of employees per establishment is growing. The authors also show that multi-unit establishments make up a relatively constant and small proportion of that sector and that business establishment births are in decline. Consolidation to exploit economies of scale in FAI is therefore likely an ongoing process that involves firms deciding to stay in business at their current location, to grow at their current location, go out of business, or move. This observation has two implications for FAI location research:

1. The classic location question from industrial recruitment literature, that is, how to attract a branch plant, takes on a slightly different form with FAI, implying a need to explore fully characteristics associated with viability of existing locations rather than focusing exclusively on decisions involving new locations.
2. A secular trend towards fewer and larger FAI establishments translates into more counties with suppressed data, leading to a need to aggregate across industries for a full data set. Aggregation can produce important shortcomings in estimations, as critical locational factors for one industry may be very different than another, even if they occupy the same two-digit NAICS sector. Thus, over time, as the number of suppressions grows, models based on public data may become less accurate.

Modeling approaches from industrial recruitment inform our analytical choices in exploring determinants of FAI firm locations. In the context of rural areas (non-FAI), manufacturing arises from several conditions. Importantly, the location choice literature reveals that rural areas can be attractive to manufacturing firms because wages, property taxes, and land costs can be lower than in metro areas, though rural workforce size can be a limitation. Unfortunately, research into rural manufacturing locational outcomes remains hamstrung by rural data suppression, which worsens with industry disaggregation.

A demand threshold is defined as the minimum market size (population) required to support a particular type of retail or service business and still yield a rate of return such that the business will continue to operate (Berry and Garrison 1958a and 1958b; Parr and Denike 2016). In examining off-farm FAI service industries, such as agricultural support services, we draw on past retail and service industry demand threshold analysis. Cleary et al. (2018) adapt Bresnahan and Reiss (1991) to estimate the minimum market size (population thresholds) that imply profitability for one or more large grocery stores in non-metro U.S. counties. The intuition is straightforward, with lower population thresholds implying store profitability in less populated areas. Their results suggest that the cost-effectiveness of broad-based policy solutions to improve physical access to large food stores or to stimulate demand may be limited. Similarly, threshold analyses could serve as a useful tool to highlight over-investment in off-farm FAI, such as may be the case with food hubs (Cleary et al. 2019). More specific FAI and FAI-related threshold analyses and more precise policy implications are now possible with the FSRDC system.

For off-farm FAI manufacturing, most research focuses on the locational determinants of food processing and food manufacturing. Typically, these studies only examine a specific geographic region, probably due to non-disclosures. Food manufacturing locations tend to occur in counties with access to input and product markets, developed

transportation networks, agglomeration economies (efficiencies that occur when industries locate near each other and industrial clusters), favorable fiscal policies, and low wages (Davis and Schluter 2005; Henderson and McNamara 1997, 2000). Particularly, important factors include proximity of agricultural commodities and low-cost labor. Rural counties tend to historically have a comparative disadvantage for attracting food processors, though non-metropolitan counties adjacent to urban areas may have an advantage for some food manufacturers (Lambert and McNamara 2009). Studies enhancing the understanding of agglomeration economies may be particularly important for rural areas, where there may be possible negative effects of competition from growing concentrations of firms (Schmit and Hall 2013). Consolidation under spatial monopsony may also result in unserved areas.

Less discussed in prior location modeling efforts is the role of industry aggregation from nesting more refined industry categories in limiting the accuracy of model outcomes. This may be a function of prior limits on alternative levels of aggregation. In contrast, aggregation bias is familiar to analysts working in international trade (Feenstra and Hanson 2000; Hillberry 2002) or regional economic models using input-output or CGE approaches (Miller and Blair 2009). Wensley and Stabler (1998) refer to this industry aggregation issue as the product mix effect. They find that their demand threshold models on subsector classifications within the NAICS system that include many more narrowly defined industries more often exhibit increasing establishment multiplication rates. Despite Wensley and Stabler's (1998) comment on the potential issue of the product mix effect from industry aggregation, most demand threshold studies count establishments at the subsector or industry group levels (Chakraborty 2012; Mushinski and Weiler 2002; Reum and Harris 2006; Shonkwiler and Harris 1996).

Clearly, there are also disadvantages in finer levels of disaggregation. No two businesses are alike; at some point of disaggregation, more output becomes cumbersome, suggesting diminishing marginal returns from disaggregation, particularly in sector classifications with small product mixes. The number of zero values across geography will also grow with greater sectoral detail. Within the FSRDC system, it becomes more feasible to test various levels of aggregation. In the analysis that follows, we explore results associated with two levels of NAICS code aggregation. The next section provides a basis for the model structure used in our analysis.

Estimation methods in the context of FAI

This section first outlines estimation methods of traditional locational choice and demand threshold literatures before weighing the trade-offs with several different estimators, largely following the recent review by Carpenter, Van Sandt, and Loveridge (2021) and applying their guidance to the FAI context.

Carlton (1983) and Bartik (1985) are among the first to model the determinants of business location using multinomial logit (McFadden 1973). Limitations of this approach include an excessive number of alternatives and violation of the independence of irrelevant alternatives (IIA) assumption, given the location alternatives are likely correlated, even when conditioning on observable characteristics (Wooldridge 2010). Researchers aggregated alternatives to address the excessive alternatives problem and attempted to control for the IIA violation by introducing spatial group indicator variables (Bartik 1985).⁶ Despite these efforts, the limitations led to use of nested logit, which uses a smaller sample

⁶For example, researchers often aggregate alternatives to the state level, arguing that U.S. state aggregations represent true alternatives considered by firms, and add regional Census or similar regional variables to control for otherwise unobserved regional correlation.

of alternatives randomly drawn from the full set of alternatives (Friedman et al. 1992; Woodward 1992). This approach faces limitations both in the imposition of an arbitrarily limited and hierarchical choice set on firms, and the implied reduction in information. Furthermore, both approaches are limited because they are only consistent if the IIA assumption holds within subsets of the alternatives, that is, various hierarchies for the nested logit and spatial groups for the multinomial logit.

Given these limitations, others model establishment location by using count data models to examine locational outcomes in a particular area (Coughlin and Segev 2000; Guimarães, Figueiredo, and Woodward 2004; Papke 1991; Wu 1999). Importantly, Guimarães, Figueiredo, and Woodward (2003, 2004) demonstrate not only that the Poisson count model presents a more tractable approach than conditional logit approaches but also that the coefficients of the Poisson model can be given an economic interpretation compatible with the framework of random utility (profit) maximization. Conveniently in the context of FAI, which include both services/retail and manufacturing, count data models are also compatible with the Central Place Theory (CPT) theoretical perspective, which researchers use in demand threshold analysis of retail and service sectors (Carpenter, Van Sandt, and Loveridge 2021). Furthermore, the count data regression is preferred when dealing with large sets of alternatives, since each spatial alternative becomes an observation, and thus, the excessive choice set problem faced by logit models becomes an advantage (Guimarães, Figueiredo, and Woodward 2003).

When modeling demand threshold and business location using count data, multiple theoretical dimensions can be considered. First, the location or “participation” decision $\{0, 1\}$ and the establishment count or “amount” decision $\{0, 1, 2, 3, \dots\}$ may be separate for FAI (and understanding each decision may be of interest), so a flexible estimator that models the decisions separately (with a zero-inflated estimator) is preferable. Relatedly, the effect of explanatory variables on the respective participation and amount decisions may have different partial effects, and there may be conditional correlation between the participation distribution and the amount distribution. Finally, both the participation and amount decision allow for the possibility of zero values. For establishment location data, there may be two types of zeros: (1) structural zeros – locations that establishments will never choose for participation; and (2) non-structural zeros – locations that establishments could potentially choose but are not chosen.

Review of corner response and count data models

Given the nature of the establishment data, $y \in \{0, 1, 2, 3, \dots\}$, count data models are most appropriate for empirical estimation.⁷ Indeed, count data estimators have become common in locational and threshold models, though researchers often fail to appropriately apply and compare the various models, as we will demonstrate. For non-FAI establishments, applications of Poisson, Negative Binomial (NB), Hurdle Poisson (HP), and Zero-Inflated Poisson (ZIP) are numerous with Carpenter, Van Sandt, and Loveridge (2021) providing a near comprehensive review for each. They also show a place for the Zero-Inflated Negative Binomial (ZINB), especially when using an expanded geographic scope.

⁷Of course, a linear model would estimate the average partial effects consistently. But problems with estimating a linear model for threshold models include that fitted establishment count values can be negative, and the assumption of a mean linear in x because $E(y|x)$ cannot truly be linear in x , with other approaches reviewed by Carpenter, Van Sandt, and Loveridge (2021)

The Poisson model remains useful at least as a starting point for comparison to other estimation procedures. However, the Poisson model's assumption that the conditional variance of the dependent variable is equal to the conditional mean is often violated in practice due to multiple potential sources of overdispersion. If the conditional distribution is overdispersed, the NB will be more efficient than Poisson. The NB allows for unobserved heterogeneity between subjects that implies overdispersion. Specifically, the NB assumes the Poisson parameter is Gamma distributed and allows the amount of overdispersion to increase with $E(y_i|x_i)$ (Wooldridge 2010).⁸ Although this likely holds when modeling FAI employment (e.g., due to rare large FAI manufacturers), this relationship is unlikely to hold when measuring FAI establishments due to excess zeros driving much of the overdispersion while *decreasing* $E(y_i|x_i)$. Hence, zero-inflated and hurdle versions of the Poisson model are often used to account for excess zeros in the data. Zero-inflated models add a logit link function which allows for an inflation stage in which the probability of an outcome being a structural zero is estimated, while still allowing for the estimation of (non-structural) zeros in the typical Poisson.⁹ This results in two sets of estimates: (1) coefficients describing the likelihood of an observation being a structural zero and (2) coefficients describing changes in the level of the dependent variable given the observation is not a structural zero. Conversely, HP only assumes one type of zero by truncating the Poisson distribution after the zero-generating process. Such a restriction would be difficult to justify for any off-farm FAI under consideration here, so we do not test the HP. See Carpenter, Van Sandt, and Loveridge (2021) for more details.

Finally, previous research failed to account for overdispersion that remains after the zero inflation in ZIP (Carpenter, Van Sandt, and Loveridge 2021). This remaining overdispersion may result from unobserved heterogeneity between subjects and excessive concentration of firms, which itself increases in economies of scale and agglomeration. Such economies are common to off-farm FAI. Thus, ZINB is underutilized in the current literature, with examinations of FAI typically using NB due to overdispersion, but failing to test for the source of that overdispersion (Bhattacharya and Innes 2016; Henderson and McNamara 2000; Lambert and McNamara 2009; Weiss and Wittkopp 2005). Indeed, we highlight that ZINB is the preferred model for numerous off-farm FAI.

Review of off-farm FAI data sources and issues

As noted above, researchers typically measure industry size in a location with establishment counts. Despite the growing body of literature highlighting effective econometric methods to measure demand thresholds on establishment counts in rural areas, extant demand thresholds research focuses on retail businesses almost exclusively and is limited in detail, or limited in industrial and geographic scope due to data limitations (Chakraborty 2012). Indeed, many demand threshold analysis and locational analysis

⁸Nonetheless, as Carpenter, Van Sandt, and Loveridge (2021) note, researchers often neglect to note that NB is not "more general" than Poisson quasi-MLE. Rather, it requires that the NB distribution be fully correct, including a specific variance/mean relationship, while the Poisson quasi-MLE is consistent for any variance assumption because it is in the linear exponential family (Gourieroux, Monfort, and Trognon 1984). Hence, Poisson quasi-MLE may be preferable, unless researchers are interested in decoupling the participation and amount marginal effects with zero-inflated models.

⁹See the Example Empirical Approach section below for the log-likelihood function.

studies use the publicly available County Business Patterns (CBP) data, resulting in several limitations¹⁰ and biased estimates (Carpenter, Van Sandt, and Loveridge 2022c).

Other federal data programs publish information about the economic activity in FAI, including the National Establishment Time-Series (NETS), Quarterly Census of Employment and Wages (QCEW), the Quarterly Workforce Indicators (QWI), and Non-employer Statistics (NS).¹¹ Although versions of the QCEW, QWI, and NS are available publicly, researchers encounter the same problem as the CBP data: the exact counts of a particular NAICS code are often suppressed. While the NETS data avoid issues with suppressions and include non-employers, it remains a smaller sample (while the LBD is essentially the universe of establishments in its frame), and thus, researchers often must pool NETS observations across multiple years to reach sufficient observations of FAI (Low et al. 2020). Isserman and Westervelt (2006) provide methods to improve over simply taking the midpoint from suppressed CBP data and numerous proprietary data sets exist that estimate suppressed cells. However, these data sets remain substantially biased, with estimates ranging from 30% attenuation bias in 3-digit NAICS county-level data to being unreliable in 5- and 6-digit NAICS data (Carpenter, Van Sandt, and Loveridge 2022c). Suppressions are particularly prevalent in some FAI due to their relatively small count in rural counties. Given the extensiveness of disclosure limitations, exact counts yield more precise models. Finally, the LBD includes all off-farm FAI establishments with paid employees without regard to whether any employees are engaged in agricultural production. An FSRDC is thus a natural place to make improvements to existing research.

Example empirical approach

In support of the objective of this paper – to demonstrate the potential value of federal administrative data for food and agricultural industry (FAI) locational outcome research – we conduct an example empirical approach. First, we compare results and model diagnostics from the Poisson, NB, and their zero-inflated counterparts to provide robust results across NAICS 115, 1151, 1152, and 1153. This flexible approach takes advantage of the count nature of the data and allows for the separate analysis of factors

¹⁰Limitations include the following: In large part due to disclosure concerns, CBP data are often suppressed, especially when examining higher digit, rural, or FAI NAICS codes; the universe of CBP businesses is only for employers. Small self-proprietary establishments with no employees – important to rural areas – are not included; in off-farm FAI, the CBP excludes farmworkers from employment measures for establishments with both farmworkers and non-farmworkers, and also excludes establishments with only farmworkers entirely; thresholds based on CBP ignore the fact that businesses of the same industry code are of different sizes. Hence, this technique could be improved if “number of establishments in an industry” are compared with more descriptive information such as “employment in an industry”; most prior studies focus on population-based estimates, which may be sensible for retail trade and other service industries, but less germane to other types of industries. For example, a meat processing plant may need local animal production more than it needs local consumers to be viable. Thus, this article also considers supply-side factors—often referred to as business location determinants in manufacturing industries—and expands current research in a way not feasible with current public data.

¹¹The CBP excludes agricultural production workers, self-employed, domestic services workers, (most) government employees, ocean-borne vessel employees, and foreign country employees. The QCEW does not exclude agricultural production workers, but has other disadvantages deriving from different reporting requirements, potentially less accurate measures of establishment counts in the context of seasonal employment, and industrial sector reclassification occurring relatively infrequently. See Becker et al. (2005) for a general discussion of differences between the sampling frames of the U.S. Census Bureau and the Bureau of Labor Statistics in the creation of these with these publicly available data. See Dunn and Hueth (2017) for a discussion specific to FAI in these data.

contributing to both structural zeros and non-structural, while also addressing the multiple potential sources of overdispersion. We selected the data generation process underlying each industry by first testing for overdispersion in the data to select between the Poisson and negative binomial distributions. After checking for zero inflation, the alpha parameter was double-checked to ensure that overdispersion was still present in the data after accounting for excess zeros.

The log-likelihood function for ZIP may be written as:

$$\ln L = \sum_{i \in S} \ln \{F(\gamma'Z_i) + [1 - F(\gamma'Z_i)] \exp(-\lambda_i)\} \\ + \sum_{i \notin S} \{\ln[1 - F(\gamma'Z_i)] - \lambda_i + y_i \beta' X_i - \ln(y_i!)\}$$

where S is a set of observations where $y_i = 0$, F is the logit link function (for zero inflation), Z is a vector containing covariates in the participation decision, and X is a vector containing covariates in the amount decision. The log-likelihood function for ZINB differs from this by allowing the Poisson parameter to be Gamma distributed thereby relaxing the assumption of equidispersion.

To choose among the various count data models empirically, much previous work has incorrectly used Vuong's Statistic to test for zero inflation (Wilson 2015).¹² To provide a consistent comparison to evaluate the various count data models, we focus on graphical distribution analysis and comparison of the Akaike and Bayesian information criteria (AIC and BIC, respectively) as suggested in Greene (1994). If an industry follows a zero-inflated data generation process, we retest the overdispersion parameter again to ensure the overdispersion was not solely a product of the zero-inflation process.

Off-Farm FAI data

We use restricted access establishment-level data from the Census' LBD and ILBD for the year 2014, along with secondary data sources to econometrically estimate the locational outcomes for the more-aggregated Support Activities for Agriculture and Forestry (NAICS 115), along with its (less-aggregated) sub-industries Support Activities for Crop Production (NAICS 1151), Support Activities for Animal Production (NAICS 1152), and Support Activities for Forestry (NAICS 1153). Table 1 provides summary statistics for these industries.

¹²The issue that previous work overlooks is that Vuong defines a model G_T to be nested in a F_Θ if and only if $G_T \subset F_\Theta$, where "c" indicates a proper subset (Vuong 1989). That is, Vuong's distributional theory refers to models that are strictly nested, strictly non-nested, and strictly overlapping models. "Strictly" in this case, implies that Vuong's results about the distribution of log likelihood ratios are not applicable in cases where models are nested at a boundary of their parameter space and when one model tends to the other when a parameter tends to infinity (Wilson 2015). These cases hold with Poisson, HP, and ZIP (and with NB and ZINB) both when using an identity link and when using a logit link for the zero-inflation parameter. Hence, the distributions are not strictly non-nested and Vuong's theory does not provide a consistent statistic for comparison. Unfortunately, the ZINB only collapses to NB when the non-negative zero-inflation parameter is at the boundary of its parameter space (i.e., equal to zero). Thus, Vuong's Statistic would not provide a consistent statistic for comparison, resulting in potentially misleading information. While providing the Vuong's statistic would conform with antecedents, providing it here would support its misuse to the detriment of this and future research in the discipline. Previous work adding Vuong's Statistic corrections (based on the Akaike and Bayesian information criteria) for bias that results from different numbers of parameters (Desmarais and Harden 2013) does not prevent the inconsistency that Wilson (2015) shows.

Table 1. Food and agricultural industry summary statistics

Industry (NAICS)	Metric	Counties	Mean	Std. Dev.	Max
Support Activities for Agriculture and Forestry (115)	Non-emp. establishments	3,093	35.59	48.20	1,016
	Emp. establishments	2,277	4.67	7.98	115
Support Activities for Crop Production (1151)	Non-emp. establishments	2,989	16.12	30.73	936
	Emp. establishments	1,506	3.10	5.55	99
Support Activities for Animal Production (1152)	Non-emp. establishments	2,949	13.40	23.28	380
	Emp. establishments	1,252	3.46	6.07	85
Support Activities for Forestry (1153)	Non-emp. establishments	2,345	3.52	6.70	67
	Emp. establishments	736	2.20	2.70	41

Note: Due to disclosure concerns, these statistics come from publicly available census county business patterns and non-employer statistics data.

Despite the inclusion of all businesses within the LBD and the ILBD, the U.S. Census Bureau requires us to conduct our analysis at an aggregated county level due to the sensitivity of the data. This unit of analysis is also necessary for including important secondary data sources that capture the potential influence of internet access, travel infrastructure, labor characteristics, urban influence, and other demographic and location variables. Further, county-level aggregation facilitates the implementation of the previously discussed methods. Table 2 provides descriptive statistics for the publicly available county-level FAI data and sources. A subset of FAI is selected to demonstrate the empirical implications of industry aggregation and size measurement in the results section.

This article does not estimate the locational outcomes of farms and ranches, but rather agricultural support services (i.e., 115, 1151, 1152, and 1153). As we describe in the next section, results vary by industry and sub-industry.

While FAI includes dozens of industry types spanning many sectors of the economy, we focus on investigating this subset of these industries with a 2014 cross section of the data using every county in the continental U.S. With the intention of highlighting the multitude of potential directions of FAI location outcomes research in the FSRDC system, we provide a diverse and large set of covariates routinely found in the location determinants literature (Carpenter, Dudensing, and Van Sandt 2022a; Carpenter et al. 2022b; Henderson, Kelly, and Taylor 2000; Lambert and McNamara 2009; Van Sandt et al. 2021a; 2021b) and to facilitate comparisons across different measures of industry size (e.g., non-employer establishments and employer establishments), covariates are the same across industries. More specifically, each model uses 2014 county-level data and includes production variables (crop farm counts, livestock farm counts, and share local food farms), place-based factors (population density and internet service providers), labor-force characteristics (social capital index, share with a bachelor’s degree or higher, share in poverty, and the unemployment rate), industrial sector interdependencies (manufacturing location quotient, and transportation and wholesale location quotient), and factor costs (property tax rate and highway density).¹³

¹³Population density is used in place of rurality measures due to convergence issues. We note that population density is highly colinear with rurality indices and related measures.

Table 2. Covariate descriptive statistics and sources

Variable	Mean	Std. Deviation	Source
Crop farms	336	313	USDA NASS
Livestock farms	327	301	USDA NASS
Local food farms	119	123	USDA NASS
Population density	0.270	1.80	ACS
Bachelor's %	13.24%	5.48	ACS
ISP count	5.19	1.12	FCC
Social capital	0.007	1.26	NERCRD
Unemployment rate (5 years. avg.)	7.89%	2.68	BLS
Manufacturing LQ	1.73	0.978	LBD
Transportation LQ	5.46	3.03	LBD
Poverty rate	16.8	6.55	ACS
Property tax rate	1.06	0.51	SA
Highway miles/sq. mi.	0.365	0.236	Census Shapefiles

Note: Table provides some descriptive statistics for model covariates and their respective sources. Local food farms are defined by USDA NASS (direct to consumer/retailer, agritourism, and value-added products). Descriptive statistics for the LQ are reproduced using publicly available U.S. Census County Business Pattern to assuage disclosure concerns. Abbreviations: ACS: American Community Survey; BLS: Bureau of Labor Statistics; FCC: Federal Communications Commission; LQ: Locational Quotient; NASS: National Agricultural Statistics Service; NERCRD: Northeast Regional Center for Rural Development at Pennsylvania State University; SA: SmartAsset.com; USDA: United States Department of Agriculture.

Example results

Tables 3 and 4 summarize the information criteria tests across support services and sub-industries for both employer establishments and non-employer establishments, respectively. The resulting distribution is itself indicative of economic and administrative features. For example, the NB distribution amongst employer establishments in Support Activities for Forestry (NAICS 1153) indicates these establishments can operate in most locations and likely serve smaller, more rural communities where market demand is not great enough to support larger employers more common in Support Activities for Crop Production (NAICS 1151) and Support Activities for Animal Production (NAICS 1152).¹⁴ Hence, industry aggregation also influences researchers' ability to identify the data generation process, with the more-aggregated Support Activities for Agriculture and Forestry (NAICS 115) preferring ZINB and obscuring this nuance. Thus, research questions specific to an industry subset of an aggregated NAICS suffer not only from estimated coefficients values being confounded by other industries' noise but also from the inability to select the industry's correct data generation process.

The marginal effects for our example industries are presented in Tables 5 and 6 to demonstrate the effect on model coefficients from (1) different levels of industry aggregation,

¹⁴It may be surprising that the information criterion indicates a slight preference to NB for Support Activities for Forestry (NAICS 1153), given there are relatively fewer counties (2345) with these types of firms present. This may result from the LBD not including public services like the forest service (but does include other pseudo-public services, like some types of hospitals).

Table 3. Information criteria for employer establishments

	Ag support (115)	Crop support (1151)	Animal support (1152)	Forest support (1153)
AIC				
Poisson	22700	17230	11730	5670
Zero-inflated Poisson	21150	15040	11000	5307
NB	15320	11650	8980	5024
Zero-inflated NB	15160	11380	8930	5006
BIC				
Poisson	22860	17380	11880	5822
Zero-inflated Poisson	21340	15230	11190	5501
NB	15470	11810	9139	5181
Zero-inflated NB	15360	11570	9132	5205

Note: Table shows the Akaike and Bayesian information criteria (AIC and BIC, respectively) across count data estimators for employer establishments in 2014. NB is negative binomial.

Table 4. Information criteria for non-employer establishments

	Ag support (115)	Crop support (1151)	Animal support (1152)	Forest support (1153)
AIC				
Poisson	63260	52170	33580	18020
Zero-inflated Poisson	62590	50450	32630	17160
NB	26460	23000	20270	14170
Zero-inflated NB	26390	22820	20150	14080
BIC				
Poisson	63420	52320	33740	18170
Zero-inflated Poisson	62790	50650	32820	17360
NB	26620	23150	20430	14330
Zero-inflated NB	26590	23020	20350	14280

Note: Table shows the Akaike and Bayesian information criteria (AIC and BIC, respectively) across count data estimators for non-employer establishments in 2014. NB is negative binomial.

which we discuss based on the difference between results for NAICS 115 compared to 1151, 1152, and 1153; and (2) different measures of industry size, which we discuss based on the difference between results for employers and non-employers. Of course, there exist unobservable variables that may affect location and correlate with explanatory variables, so

Table 5. Employer establishment locational determinants results

	Ag support (115)	Crop support (1151)	Animal support (1152)	Forest support (1153)
Inflation stage				
Crop farms	-0.001***	-0.003***	-0.001*	
	0.000	0.000	0.001	
Livestock farms	0.000	0.001***	-0.001	
	0.000	0.000	0.000	
Local food farms	-0.047**	0.057	-0.156	
	0.021	0.036	0.102	
Manufacturing LQ	0.002	0.020***	-0.009	
	0.005	0.005	0.022	
Transportation LQ	-0.000	-0.012	0.021***	
	0.004	0.008	0.006	
Highway miles	-1.521	-6.286*	-2.058	
	1.372	3.616	4.576	
Amount stage				
Crop farms	0.030***	0.026***	0.004***	0.000
	0.006	0.005	0.001	0.000
Livestock farms	-0.011***	-0.018***	0.004***	0.001***
	0.004	0.004	0.001	0.000
Local food farms	-5.061***	-6.893***	0.171	0.856***
	1.224	1.650	0.358	0.121
Population density	0.000	0.000***	-0.000	0.000
	0.000	0.000	0.000	0.000
Bachelor's %	0.625***	0.162***	0.222***	0.039***
	0.102	0.047	0.025	0.006
ISP count	0.392**	0.308**	0.024	-0.001
	0.172	0.125	0.066	0.025
Social capital	-0.529***	-0.386**	-0.447***	0.069**
	0.187	0.153	0.092	0.029
Unemployment rate	0.175	0.052	-0.170***	0.049***
	0.141	0.099	0.046	0.014
Manufacturing LQ	-0.174*	-0.224**	0.042	0.029**
	0.104	0.088	0.036	0.013

(Continued)

Table 5. (Continued)

	Ag support (115)	Crop support (1151)	Animal support (1152)	Forest support (1153)
Transportation LQ	-0.252*	-0.258**	-0.020	-0.007
	0.139	0.117	0.062	0.017
Poverty rate	0.091**	0.118***	-0.039***	0.013***
	0.0429	0.044	0.013	0.005
Property tax rate	2.920***	1.835***	0.830***	-0.106*
	0.573	0.490	0.157	0.062
Highway miles	1.822	-515.6***	212.90***	-32.70***
	120.8	150.7	45.970	11.580
Alpha	0.546***	0.719***	0.723***	1.314***
Log pseudolikelihood	-7493	-5654	-4433	-2470

Note: Table shows the results across Ag support services and its subsectors when using county-level employer establishment counts as unit of analysis. Forest support (NAICS 1153) shows no inflation stage because the information criteria testing preferred negative binomial to zero-inflated negative binomial (see Table 5).

Abbreviations: Bachelors %, percent of the population with a bachelor's degree or higher; ISP, internet service provider; LQ, location quotient.

Significance levels: *** $p < 1\%$, ** $p < 5\%$, * $p < 10\%$.

we again emphasize the associative and non-causal interpretation of these example results, while focusing on (1) and (2).

In all zero-inflated models, the logistic function predicts whether a county is a structural zero. Thus, the negative coefficients in the inflation stage of Tables 5 and 6 indicate that an increase in the variable is expected to lead to a decrease in the likelihood of the county being a structural zero. Following a priori expectations, the presence of agricultural production businesses corresponds to a smaller likelihood of a county being a structural zero for 115, 1151, and 1152 for both employer and non-employer establishments. Inflation stage results also show a positive association between above-average manufacturing industry concentration and a county having zero crop support services (1151) structurally, and a positive association between transportation industry concentration and a county having animal support services (1152) structurally. Researchers examining the more-aggregated agricultural support services (115) would find statistically insignificant coefficients for both manufacturing and transportation location quotients.

To examine the effects of different levels of industry aggregation, in the agricultural and forestry support services industries, we compare the marginal effects in the amount stage presented in Tables 5 and 6 across the more-aggregated NAICS 115 with its three subsectors. Results suggest that aggregating industries can influence the magnitude, direction, and statistical significance of locational factors for establishments involved in support activities for crop and animal production. This finding is clear when examining the association of crop farms and livestock farms with the location of support services establishments. For example, Table 5 shows that local livestock producers have a negative relationship with the more-aggregated Support Activities for Agriculture and Forestry (NAICS 115), but that this is driven by the relationship with Support Activities for Crop Production (1151), while the relationship with Animal Support Services (1152)

Table 6. Non-employer establishment locational determinants results

	Ag support (115)	Crop support (1151)	Animal support (1152)	Forest support (1153)
Inflation stage				
Crop farms	-0.001*	-0.001***	-0.000	-0.000
	0.000	0.000	0.000	0.000
Livestock farms	-0.001*	0.000	-0.001***	-0.000
	0.000	0.000	0.000	0.000
Local food farms	0.007*	0.003	0.014	-0.011
	0.004	0.011	0.010	0.024
Manufacturing LQ	0.000	0.001	0.003	-0.172***
	0.001	0.002	0.002	0.023
Transportation LQ	0.001	-0.002	0.002	0.006***
	0.001	0.003	0.003	0.002
Highway miles	-0.850**	-3.004***	-2.958***	-2.428
	0.352	0.981	0.749	1.656
Amount stage				
Crop farms	0.045**	0.119***	0.018***	0.002**
	0.023	0.027	0.003	0.001
Livestock farms	0.065***	-0.038***	0.042***	0.007***
	0.016	0.014	0.004	0.001
Local food farms	0.156***	-18.57***	9.542***	5.362***
	0.048	4.849	1.598	0.525
Population density	0.002	0.001	0.000	0.000
	0.001	0.001	0.000	0.000
Bachelor's %	2.054***	1.217***	0.798***	0.200***
	0.505	0.306	0.077	0.025
ISP count	-0.436	-0.681	-0.195	-0.369***
	1.406	1.058	0.254	0.096
Social capital	-6.694***	-5.565***	-2.359***	-0.421***
	1.677	1.266	0.353	0.091
Unemployment rate	-1.769	-1.592**	-0.926***	0.206***
	1.124	0.810	0.168	0.073
Manufacturing LQ	0.172	-0.513	0.148	0.490***
	0.386	0.336	0.131	0.085

(Continued)

Table 6. (Continued)

	Ag support (115)	Crop support (1151)	Animal support (1152)	Forest support (1153)
Transportation LQ	-1.126*	-1.335**	-0.227	-0.056
	0.680	0.537	0.169	0.061
Poverty rate	0.254	0.615***	-0.14***	0.034**
	0.220	0.219	0.052	0.018
Property tax rate	9.938***	4.163**	3.488***	0.627***
	2.904	2.120	0.617	0.229
Highway miles	1367**	437.6	650.8***	-15.03
	686.0	509.6	183.9	47.12
Alpha	0.356***	0.520***	0.369***	0.592***
Log pseudolikelihood	-13160	-11380	-10040	-7009

Note: Table shows the results across Ag support services and its subsectors when using county-level non-employer establishment counts as unit of analysis.

Abbreviations: Bachelors %, percent of the population with a bachelor's degree or higher; ISP, internet service provider; LQ, location quotient.

Significance levels: *** $p < 1\%$, ** $p < 5\%$, * $p < 10\%$.

and Forestry Support Services (1153) reverses direction and is positive. The relationship between crop farms and our selection of industries is positive, but varies in statistical significance and is statistically insignificant for Forestry Support Services. This statistically significant variation and sign reversal occurs across a number of variables common to location choice research, including the share of the population with a bachelor's degree, social capital, unemployment rate, industrial location quotients, poverty rate, property tax rate, and highway miles. To summarize, using industry classifications that are more broadly defined than implied by a research question leads to statistically significant over- or underestimating the influence of important location factors specific to a sub-industry of interest.

To examine the effect of using different measures of industry size, we compare marginal effects estimated with employer establishments (Table 5) to marginal effects with non-employer establishments (Table 6). The differences between these associations are statistically significant for multiple covariates. For example, livestock farms have a negative association with the number of employer businesses in the aggregated 115, but a large positive association with the number of 115 non-employer businesses. (The sub-industries are more stable directionally.) The association between the local poverty rate and the aggregated 115 loses statistical significance for non-employers, while the sub-industries maintain significance and directional consistency with employer establishments. More generally, the less-aggregated industries show a much higher share of significant covariates in the amount stage of the models, indicating that more specific industry codes may benefit from less noise resulting from the mix of competing industry-specific nuances in aggregated 115 classifications. In summary, it is clear from the amount stage results that non-employer FAI establishments are different from the locational determinants than employer FAI.

Summary and concluding remarks

An increasingly large share of the food and fiber supply chain has moved beyond the farm gate, leading to FAI being located across different sectors of the economy. Despite this shift, research aimed at selecting factors associated with locational outcomes of these FAI remains difficult, due to the rural nature of FAI and the associated suppression of FAI metrics in publicly available data sets. Previous demand threshold literature circumnavigates this issue by limiting the geographic scope of their study area, aggregating industries to more general classifications, and by adopting econometric methods that may fit and describe the sample but not the population of interest.

Our review and supporting example results using FSRDC data demonstrate the potential value of using FSRDC data to conduct FAI locational analysis at various levels of aggregation. In particular, we show that the choice of industrial aggregation and size measurement often affect the magnitude, sign, and significance of several variables common to locational choice research. Hence, researchers using aggregated FAI industries should take care when generalizing their results (to all sub-industries). Our supporting results make contributions to the literature by employing zero-inflation count data models to restricted access data sets through the Federal Statistical Research Data Center system to estimate demand threshold models for several FAI. We find that while the ZINB data generation process is almost always preferred across FAI. The distributions of these industries reveal important findings: (1) the significant overdispersion present in all FAI signifies the tendency of FAI to cluster; and (2) the variety of agricultural support services have different data generation processes with nuanced structural and sampling zeros.

Estimated coefficients across employers and non-employers show the benefits of more finely disaggregated industry classifications. Notably, in agricultural support services, broader industry classifications lead to underestimation or even sign reversal of statistically significant location factors. The importance of non-employer establishments to rural economies and their different locational determinants compared to employer establishments is also evident through differences in the models' assumed distributions, the magnitudes of coefficient marginal effects, and the differences in significant coefficients.

Some data limitations persist, even in an FSRDC. For example, researchers cannot comingle USDA NASS (National Agricultural Statistics Service) Census of Agriculture microdata in the FSRDC. Thus, integration of the two data sources requires either taking (suppressed and aggregated) public versions of FSRDC data into the USDA NASS laboratories or taking (suppressed and aggregated) public versions of the NASS data into an FSRDC laboratory. Although the methods section highlights the benefits of using aggregated count data in the context of FAI locational outcomes, future research opportunities would further expand, given the ability to comingle USDA NASS and FSRDC microdata. This comingled data would not only facilitate substantial research on the interplay between on- and off-farm business dynamics over time and with respect to policy (and structural) changes but also improve the seasonal measures of industry size in FAI.

The USDA remains an outlier in their provision of microdata to FSRDC researchers, as current federal entities that allow comingling of data in the FSRDC system (directly or indirectly) include the U.S. Census Bureau, Internal Revenue Service, Bureau of Labor Statistics, Bureau of Justice Statistics, Environmental Protection Agency, National Center for Health Statistics, and Agency for Healthcare Research and Quality, among several others. This policy by USDA is needlessly limiting the research questions researchers can answer with USDA data. Our example results highlight that compelling researchers to use relatively aggregated industry classifications can lead to entirely incorrect conclusions when attempting to generalize to FAI subsectors.

Despite the contributions made here, there are still many research questions that cannot be answered with current public data resources. These questions are not only central to agricultural marketing and economic development, but the livelihoods of the people who depend on FAI as well as other frequently censored industries primarily found in rural areas. With this in mind, we highlight three areas using FSRDC data resources for future research. First, while specific industrial quality comparisons between public and restricted access data are understandably prohibited within the FSRDC, comparisons between restricted access and purchased estimates may reveal the true value of these expensive yet proprietary data sets beyond the broad implications of Carpenter, Van Sandt, and Loveridge (2022).¹⁵ Second, aggregation bias may vary across space. Western counties are larger than eastern counties, and this may lead to less suppression and aggregation bias. Third, modeling more disaggregated industries within FAI would improve our knowledge of agricultural supply chain location decisions and their impact on rural economies. Finally, the different industry size measures used here could be expanded beyond FAI to explore the data generation and decision processes of other industries and sub-industries, particularly those that may be germane to rural economic development that are often suppressed in public data.

Data availability statement. The data that support the findings for this article are available through the Federal Statistical Research Data Center (FSRDC). For details on how to access the FSRDC system, visit <https://www.census.gov/about/adrm/fsrdc.html>. Other data are publicly available, with details on sources available in Table 2.

Acknowledgements. The authors thank Rebekka Dudensing and numerous conference participants for their comments on an earlier version of this work. Any opinions and conclusions expressed herein are those of the author and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

Author contributions. CC, AV, and SL conceived and designed the study. AV conducted data gathering. CC and AV performed statistical analyses. CC, AV, and SL wrote the article.

Funding statement. The project was supported by the Agricultural and Food Research Initiative Competitive Program of the USDA National Institute of Food and Agriculture (NIFA), award numbers 2017-67023-26242 and 2018-68006-27641.

Competing interests. The authors declare none.

References

- Ahearn, M.C., J. Yee, and P. Korb 2005. "Effects of differing farm policies on farm structure and dynamics." *American Journal of Agricultural Economics* 87 (5): 1182–1189.
- Akobundu, E., J. Alwang, A. Essel, G.W. Norton, and A. Tegene 2004. "Does extension work? Impacts of a program to assist limited-resource farmers in Virginia." *Applied Economic Perspectives and Policy* 26 (3): 361–372.
- Arauzo-Carod, J.-M., and E. Viladecans-Marsal 2009. "Industrial location at the intra-metropolitan level: the role of agglomeration economies." *Regional Studies* 43 (4): 545–558.
- Bartik, T.J. 1985. "Business location decisions in the United States: estimates of the effects of unionization, taxes, and other characteristics of states." *Journal of Business & Economic Statistics* 3 (1): 14–22.

¹⁵There are exceptions in public data availability. For example, forest economists examine mill locational choice as a function of the wood within a certain area and have developed their own wood and mill locational data, in part using the Forest Service database.

- Becker, R., J. Elvery, L. Foster, C.J. Krizan, S. Nguyen, and D. Talan 2005. "A Comparison of the business registers used by the bureau of labor statistics and the bureau of the census." In *Proceedings of the American Statistical Association, Business and Economics Section*, Alexandria, VA.
- Berry, B.J.L., and W.L. Garrison 1958a. "A note on central place theory and the range of a good." *Economic Geography* 34 (4): 304.
- Berry, B.J.L., and W.L. Garrison 1958b. "Recent developments of central place theory." *Papers in Regional Science* 4 (1): 107–120.
- Bhattacharya H., and R. Innes 2016. "Concentration, product variety, and entry-for-merger: Evidence from new product introductions in the U.S. food industry." *American Journal of Agricultural Economics* 98 (5): 1360–1376.
- Bresnahan, T.F., and P.C. Reiss 1991. "Entry and competition in concentrated markets." *Journal of Political Economy* 99 (5): 977–1009.
- Carlton, D.W. 1983. "The location and employment choices of new firms: An econometric model with discrete and continuous endogenous variables." *The Review of Economics and Statistics* 65 (3): 440.
- Carpenter, C. W., A.T. Van Sandt, R. Dudensing, and S. Loveridge 2022b. "Profit pools and determinants of potential county-level manufacturing growth." *International Regional Science Review* 45 (3): 188–224.
- Carpenter, C.W., R.M. Dudensing, and A.T. Van Sandt 2022a. "Estimating determinants of transportation and warehousing establishment locations using U.S. administrative data." *REGION* 9 (1): 1–27.
- Carpenter, C.W., and S. Loveridge 2018. "Differences between Latino-owned businesses and white-, black-, or Asian-owned businesses: Evidence from census microdata." *Economic Development Quarterly* 32 (3): 225–241.
- Carpenter, C.W., and S. Loveridge 2019. "A spatial model of growth relationships and Latino-owned business." *The Annals of Regional Science* 63 (3): 541–557.
- Carpenter, C.W., and S. Loveridge 2020. "Business, Owner, and regional characteristics in latino-owned business growth: An empirical analysis using confidential census microdata." *International Regional Science Review* 43 (3): 254–285.
- Carpenter, C.W., A.T. Van Sandt, and S. Loveridge 2021. "Empirical methods in business location research." *Regional Studies, Regional Science* 8 (1): 344–361.
- Carpenter, C.W., A.T. Van Sandt, and S. Loveridge 2022c. "Measurement error in U.S. regional economic data." *Journal of Regional Science* 62 (1): 57–80.
- Chakraborty, K. 2012. "Estimation of minimum market threshold for retail commercial sectors." *International Advances in Economic Research* 18 (3): 271–286.
- Chen, M.X., and M.O. Moore 2010. "Location decision of heterogeneous multinational firms." *Journal of International Economics* 80 (2): 188–199.
- Cleary, R., A. Bonanno, L. Chenarides, and S.J. Goetz 2018. "Store profitability and public policies to improve food access in non-metro U.S. counties." *Food Policy* 75: 158–170.
- Cleary, R., S.J. Goetz, D.T. Mcfadden, and H. Ge 2019. "Excess competition among food hubs." *Journal of Agricultural and Resource Economics* 44 (1): 141–163.
- Coughlin, C.C., and E. Segev 2000. "Location determinants of new foreign-owned manufacturing plants." *Journal of Regional Science* 40 (2): 323–351.
- Davis, D.E., and G.E. Schluter 2005. "Labor-force heterogeneity as a source of agglomeration economies in an empirical analysis of county-level determinants of food plant entry." *Journal of Agricultural and Resource Economics* 30 (3): 480–501.
- Davis, S.J., R.J. Faberman, and J. Haltiwanger 2006. "The flow approach to labor markets: New data sources and micro–macro links." *Journal of Economic Perspectives* 20 (3): 3–26.
- Desmarais, B.A., and J.J. Harden 2013. "Testing for zero inflation in count models: Bias correction for the Vuong test." *The Stata Journal* 13 (4): 810–835.
- Devereux, M.P., R. Griffith, and H. Simpson 2007. "Firm location decisions, regional grants and agglomeration externalities." *Journal of Public Economics* 91 (3–4): 413–435.
- Dunn, R.A., and B. Hueth 2017. "Food and agricultural industries: Opportunities for improving measurement and reporting." *American Journal of Agricultural Economics* 99 (2): 510–523.
- Feenstra, R.C., and G.H. Hanson 2000. "Aggregation bias in the factor content of trade: Evidence from U.S. manufacturing." *American Economic Review* 90 (2): 155–160.
- Figureiredo, O., P. Guimarães, and D. Woodward 2002. "Home-field advantage: Location decisions of Portuguese entrepreneurs." *Journal of Urban Economics* 52 (2): 341–361.

- Foster, L., J. Haltiwanger, and C.J. Krizan 2006. "Market selection, reallocation, and restructuring in the U.S. retail trade sector in the 1990s." *Review of Economics and Statistics* 88 (4): 748–758.
- Friedman, J., D.A. Gerlowski, and J. Silberman 1992. "What attracts foreign multinomial corporations? Evidence from branch plant location in the United States." *Journal of Regional Science* 32 (4): 403–418.
- Goetz, S.J., and M. Davlasheridze 2017. "State-level cooperative extension spending and farmer exits." *Applied Economic Perspectives and Policy* 39 (1): 65–86.
- Gourieroux, C., A. Monfort, and A. Trognon 1984. "Pseudo maximum likelihood methods: Applications to poisson models." *Econometrica* 53 (3): 701–720.
- Greene, W.H. 1994. "Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models." NYU Working Paper No. EC-94-10.
- Guimarães, P., O. Figueiredo, and D. Woodward 2003. "A tractable approach to the firm location decision problem." *Review of Economics and Statistics* 85 (1): 201–204.
- Guimarães, P., O. Figueiredo, and D. Woodward 2004. "Industrial location modeling: extending the random utility framework." *Journal of Regional Science* 44 (1): 1–20.
- Haltiwanger, J., R.S. Jarmin, and J. Miranda 2013. "Who creates jobs? Small versus large versus young." *Review of Economics and Statistics* 90 (2): 347–361.
- Henderson, J.R., and K.T. McNamara 1997. "Community attributes influencing local food processing growth in the U.S. Corn Belt." *Canadian Journal of Agricultural Economics/Revue Canadienne D'Agroeconomie* 45 (3): 235–250.
- Henderson, J.R., and K.T. McNamara 2000. "The location of food manufacturing plant investments in Corn Belt counties." *Journal of Agricultural and Resource Economics* 25 (2): 680–697.
- Henderson, J.W., T.M. Kelly, and B.A. Taylor 2000. "The impact of agglomeration economies on estimated demand thresholds: An extension of Wensley and Stabler." *Journal of Regional Science* 40 (4): 719–733.
- Hertz, T., and S. Zahniser 2013. "Is there a farm labor shortage?" *American Journal of Agricultural Economics* 95 (2): 476–481.
- Hillberry, R.H. 2002. "Aggregation bias, compositional change, and the border effect." *Canadian Journal of Economics/Revue Canadienne D'Economie* 35 (3): 517–530.
- Holl, A. 2004. "Transport infrastructure, agglomeration economies, and firm birth: Empirical evidence from Portugal." *Journal of Regional Science* 44 (4): 693–712.
- Isserman, A.M., and J. Westervelt 2006. "1.5 Million missing numbers: Overcoming employment suppression in county business patterns data." *International Regional Science Review* 29 (3): 311–335.
- Jarmin, R.S., and J. Miranda 2002. *The Longitudinal Business Database* (No. 02–17).
- Lambert, D.M., and K.T. McNamara 2009. "Location determinants of food manufacturers in the United States, 2000–2004: Are nonmetropolitan counties competitive?" *Agricultural Economics* 40 (6): 617–630.
- Low, S.A., M. Bass, D. Thilmany, and M. Castillo 2020. "Local foods go downstream: Exploring the spatial factors driving U.S. food manufacturing." *Applied Economic Perspectives and Policy* 43 (3): 896–915.
- McFadden, D. 1973. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, ed., *Frontiers in Econometrics*. New York, NY: Academic Press, pp. 105–142.
- Miller, R.E., and P.D. Blair 2009. *Input-Output Analysis Foundations and Extensions*, 2nd edn. Cambridge: Cambridge University Press.
- Mushinski, D., and S.S. Weiler 2002. "A note on the geographic interdependencies of retail market areas." *Journal of Regional Science* 42 (1): 75–86.
- Papke, L.E. 1991. "Interstate business tax differentials and new firm location: Evidence from panel data." *Journal of Public Economics* 45 (1): 47–68.
- Parr, J.B., and K.G. Denike 2016. "Theoretical problems in central place analysis." *Economic Geography* 46 (4): 568–586.
- Reum, A.D., and T.R. Harris 2006. "Exploring firm location beyond simple growth models: A double hurdle application." *The Journal of Regional Analysis & Policy* 36 (1): 45–67.
- Schmit, T.M., and J.S. Hall 2013. "Implications of agglomeration economies and market access for firm growth in food manufacturing." *Agribusiness* 29 (3): 306–324.
- Shonkwiler, J.S., and T.R. Harris 1996. "Rural retail business thresholds and interdependencies." *Journal of Regional Science* 36 (4): 617–630.

- Van Sandt, A., C.W. Carpenter, R. Dudensing, and S. Loveridge** 2021a. "Estimating determinants of healthcare establishment locations with restricted federal administrative data." *Health Economics* **30** (6): 1328–1346.
- Van Sandt, A., C.W. Carpenter, R.M. Dudensing, and S. Loveridge** 2021b. "Estimating determinants of health care establishment locations with restricted federal administrative data." *Health Economics* **30** (6): 1328–1346.
- Vuong, Q.H.** 1989. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica* **57** (2): 307.
- Weiss, C.R., and A. Wittkopp** 2005. "Retailer concentration and product innovation in food manufacturing." *European Review of Agricultural Economics* **32** (2): 219–244.
- Wensley, M.R.D., and J.C. Stabler** 1998. "Demand-threshold estimation for business activities in rural Saskatchewan." *Journal of Regional Science* **38** (1): 155–177.
- Wilson, P.** 2015. "The misuse of the Vuong test for non-nested models to test for zero-inflation." *Economics Letters*, **127**: 51–53.
- Woodward, D.P.** 1992. "Locational determinants of Japanese manufacturing start-ups in the United States." *Southern Economic Journal* **58** (3): 690.
- Wooldridge, J.M.** 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge, MA: MIT Press.
- Wu, F.** 1999. "Intrametropolitan FDI firm location in Guangzhou, China." *The Annals of Regional Science* **33** (4): 535–555.