Food flows between counties in the United States

Environmental Research Letters

LETTER

Food flows between counties in the United States

Xiaowen Lin[1], Paul J Ruess[1], Landon Marston[2] and Megan Konar[1,3]

[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana IL, United States of America
[2] Department of Civil and Environmental Engineering, Kansas State University, Manhattan, KS, United States of America
[3] Author to whom any correspondence should be addressed.

E-mail: mkonar@illinois.edu

Keywords: food flows, networks, algorithm development

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
23 August 2018

REVISED
11 June 2019

ACCEPTED FOR PUBLICATION
13 June 2019

PUBLISHED
26 July 2019

Abstract

Food consumption and production are separated in space through flows of food along complex supply chains. These food supply chains are critical to our food security, making it important to evaluate them. However, detailed spatial information on food flows within countries is rare. The goal of this paper is to estimate food flows between all county pairs within the United States. To do this, we develop the Food Flow Model, a data-driven methodology to estimate spatially explicit food flows. The Food Flow Model integrates machine learning, network properties, production and consumption statistics, mass balance constraints, and linear programming. Specifically, we downscale empirical information on food flows between 132 Freight Analysis Framework locations (17 292 potential links) to the 3142 counties and county-equivalents of the United States (9869 022 potential links). Subnational food flow estimates can be used in future work to improve our understanding of vulnerabilities within a national food supply chain, determine critical infrastructures, and enable spatially detailed footprint assessments.

1. Introduction

Most food security research focuses on increasing production (Lobell *et al* 2011, Long *et al* 2015, Liang *et al* 2017), but distribution through complex supply chains is also critical to food security (Ercsey-Ravasz *et al* 2012, Konar *et al* 2018). Food supply chains are increasingly complex and global in scope, incorporating the production, distribution, and consumption of food commodities (Porkka *et al* 2013, MacDonald *et al* 2015). Here, we refer to the movement of food through complex supply chains within a country as 'food flows', reserving the term 'food trade' for the international trade of food commodities. Food flow networks depend on many factors, such as production locations, population centers, storage and transport infrastructure, and socio-political factors (Venkatramanan *et al* 2017). It is increasingly important to evaluate food flow networks, since these coupled human-natural systems can have dramatic implications for the environment and underpin our food security (Dalin and Rodriguez-Iturbe 2016, Seekell *et al* 2017). Spatially detailed food flow estimates would improve our understanding of food supply chain vulnerabilities and enable spatially detailed footprint assessments. However, we know relatively little about food flows due to a sparsity of data, with the exception of international food trade. The goal of this paper is to estimate food flows between all county pairs within the United States.

The United States is a key country in the global food system (Xu *et al* 2011). The US produces over 30% of the world's corn and over 50% of the world's soybeans (USDA 2013). The US also accounts for large shares of the world export market for several staples: about 60% for corn, 40% for soybeans, 25% for wheat, and 70% for sorghum (USDA 2013), making the US an important contributor to global grain supplies (FAO 2013). The ability to grow and transport agricultural products enables the US to provide both domestic and global food security (Lin *et al* 2014). The US is able to maintain its role as a key agricultural producer, consumer, and trade power largely due to its supporting institutions (e.g. agricultural subsidies, crop insurance, etc) and infrastructure (e.g. irrigation systems, food distribution infrastructure, etc) (Deryugina and Konar 2017, Marston *et al* 2018, Rushforth and Ruddell 2018). Supply chains in the US

Figure 1. Maps of political boundaries within the United States. (A) Map of FAF zones. (B) Map of the counties of the United States.

are also responsible for a large national carbon (Weber and Matthews 2008, Cuéllar and Webber 2010, Liang *et al* 2016), water (Dang *et al* 2015, Vora *et al* 2017, Wang *et al* 2017), and chemical pollution footprint (Nesheim *et al* 2015).

Data on subnational food flows is available within the United States at a coarse spatial resolution. This availability of subnational food flow information is a major reason for our selection of the US for this work. The US Census Bureau and the Bureau of Transportation Statistics produce the Commodity Flow Survey (CFS) every five years (ending in '2' and '7'). The CFS provides data on the movement of commodities in the United States, including their value, weight, and mode of transportation, as well as the origin and destination of shipments from manufacturing, mining, wholesale, and selected retail and services establishments. The Freight Analysis Framework (FAF) builds on the CFS data to provide data on freight movement between the 132 FAF zones of the US (see figure 1(A)) (Oak Ridge National Laboratory 2015). FAF reports flows of coarse food commodity classes (see table 1). This census information on food flows within the US has been used to evaluate their vulnerabilities at a relatively coarse spatial scale (Lin *et al* 2014). Spatially refined food flows would enable future research to better understand the potential vulnerabilities and resiliencies within the US food supply chain, and would advance lifecycle and footprint assessments.

Table 1. List of standard classification of transported goods (SCTG) food categories included in this study.

SCTG	Model
1	Animals and fish (live)
2	Cereal grains (includes seed)
3	Agricultural products (excludes animal feed, cereal grains, and forage products)
4	Animal feed, eggs, honey, and other products of animal origin
5	Meat, poultry, fish, seafood, and their preparations
6	Milled grain products and preparations, and bakery products
7	Other prepared foodstuffs, fats and oils

Our work contributes to the recent literature that models food flows. A few recent papers have modeled spatially detailed food flows (Smith *et al* 2017, Venkatramanan *et al* 2017). Venkatramanan *et al* (2017) present a data-driven approach to estimate food flows between markets in Nepal in order to evaluate their propensity to spread pests. Smith *et al* (2017) use a transportation optimization model to estimate corn flows between US counties. Our approach is related but distinct. As in the existing literature, we use food production and consumption statistics in conjunction with a linear programming framework that minimizes transport distance. In this paper, we add some novel elements to the food modeling literature. First, we constrain our food flows to have the same

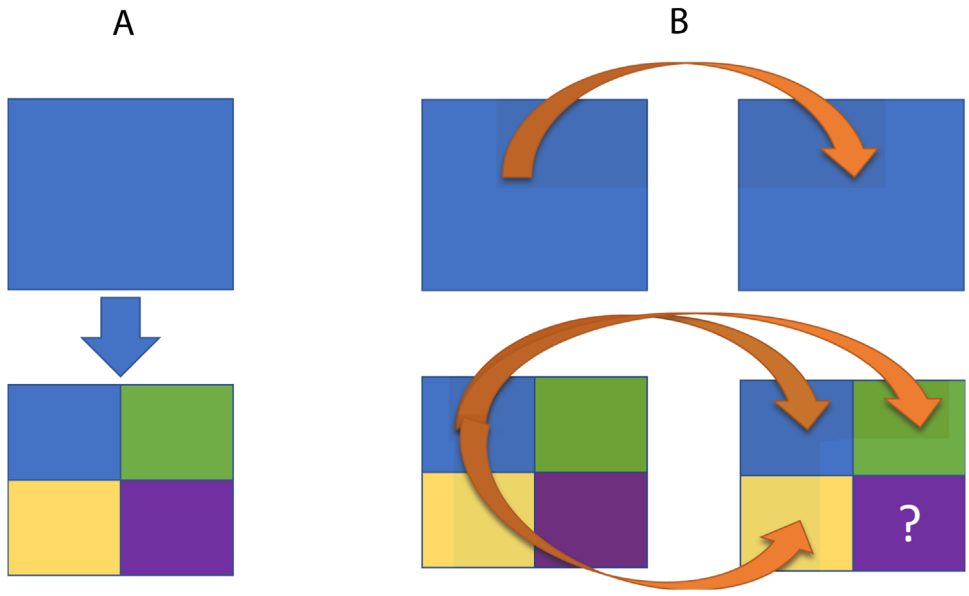

Figure 2. Conceptual schematic of the problem statement. The goal of the Food Flow Model is to downscale spatial flow data. This is more complex than traditional spatial downscaling (A), which typically aims to estimate attributes of sub-entities within a larger entity. Now, we aim to estimate connections and mass transfers between all node pairs (B). There are many more node pairs than spatial units, because the number of node pairs (i.e. links) scales as $n^*(n-1)$, where n is the number of spatial units.

network properties as those of the observed food flow networks (Konar *et al* 2018). Recent research has shown that global, subnational, and village scale food flow networks share structural properties, including a gamma mass flux distribution (Konar *et al* 2018). Second, we incorporate the principle of mass balance in our model. We do this by requiring that the food flows from counties within an FAF zone sum to the food flows from that FAF zone. Both of these novel aspects enhance the realism of our approach.

The goal of this paper is to estimate food flows between all counties and county equivalents in the United States. There are 3142 counties and county equivalents in the United States: 3007 counties, 64 parishes, 19 organized boroughs, 10 census areas, 41 independent cities, and the District of Columbia. For the rest of this paper, we refer to these 'counties and county equivalents' simply as 'counties' for short. The major question that we address is: what are the food flows between counties within the United States? To answer this question, we develop the Food Flow Model, a novel, data-driven framework to estimate food flows in locations without empirical data. We detail our methods in section 2. We discuss our results in section 3. We conclude in section 4.

2. Methods

We develop a novel methodology to estimate food transfers between counties in the United States. To do this, we downscale data on food transfers at the FAF zone spatial scale (refer to figure 1(A)) to counties within the United States (refer to figure 1(B)). From figure 1 it is clear that our goal requires the estimation of flows at a much finer spatial resolution (i.e. between all 3142 county pairs) than that for which information is available (i.e. between 132 FAF zones). Since the number of directed paths is determined by $(n)(n-1)$, this means that our goal requires that we move from a system with 17 292 potential links ($n = 132$ at FAF zone scale) to estimating 9869 022 potential links ($n = 3142$ at county scale). As such, flow estimation quickly increases in complexity and computational demands as the number of nodes increases. In this way, our problem is distinct to most other spatial downscaling problems, in which a coarse spatial value is assigned to entities within its domain (see figure 2(A)). Instead, we want to downscale *flows*, which requires estimating values (including zeros) between all node pairs (i.e. links) in our system. Figure 2 presents a conceptual framing of this challenge.

To achieve our goal, we develop the Food Flow Model, a computational algorithm that integrates machine learning, linear programming, network constraints, and mass balance (see figure 3 for a schematic of our modeling approach). We incorporate known properties of food flow networks at different spatial scales (Konar *et al* 2018) through the development of a gamma mixture hurdle model. This approach ensures that estimated mass fluxes follow a gamma distribution as in empirical networks (e.g. see Konar *et al* 2018). We use supervised learning (a machine learning technique) to establish a gamma mixture hurdle model at the FAF spatial scale and then use it to estimate food flow potentials between counties. This approach incorporates statistical information on crop, livestock, and other economic factors of production at

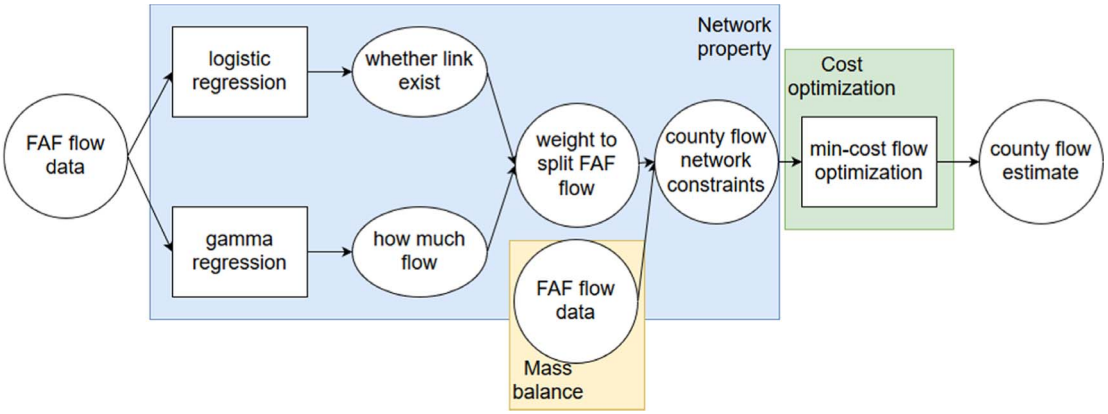

Figure 3. Schematic of the Food Flow Model.

the county spatial scale. In this way, our approach maintains realism, since food transfers are assigned to links only if the food is produced and consumed in those locations. Note that here 'consumption' refers to commodity transformation such as food processing, livestock feed, biofuel conversion, or another intermediate step in the supply chain of different goods, which means that both intermediate and final consumption are included. Then, we use data on food transfers at the FAF spatial resolution to constrain our county scale estimates. This provides a mass balance constraint to our approach. Finally, the Food Flow Model incorporates linear programming to solve the system through the minimization of the transport distance between counties. We detail our data sources and algorithm below.

2.1. Input data

We obtain two major types of data for this study. First, we obtain data on agricultural and food commodity transfers between FAF zones in the United States. The Freight Analysis Framework Version 4 (FAF4) database provides empirical agricultural and food commodity transfers between FAF zones for the year 2012 (Oak Ridge National Laboratory 2015). Second, we obtain statistical information on economic production within each US county. We obtain county-level production data for the year 2012 to match FAF4. All data sources are detailed in table 2.

The FAF4 dataset utilizes data from numerous sources to provide an exhaustive description of subnational freight movement in the United States, as well as trade with major international regions. The CFS is foundational to the FAF4 dataset. Every five years (years ending in '2' and '7'), the CFS samples more than 100,000 establishments that ship freight domestically. Survey responses are aggregated to the corresponding FAF zone, commodity class, and across the 4 quarterly surveys administered during the year of record to protect the confidentiality of survey respondents. Freight shipments within the CFS and FAF dataset are grouped into 42 classes using the two-digit standard classification of transported goods (SCTG). Here, we are primarily interested in agriculture and food goods, which are represented by SCTG 01-07 (table 1). We use the FAF4 commodity transfer database as training data for a supervised learning algorithm to determine the functional form of regression models of food transfers between FAF zones that are then applied to the county spatial scale (see the following section for more details). This approach assumes that the regression model is consistent across spatial scales. The FAF4 data is also used to constrain transfers within our county-to-county Food Flow Model. We utilize the principle of mass balance to ensure that counties located within an FAF zone do not exceed the mass flux reported at the FAF spatial scale.

Distance between all county pairs was obtained from Oak Ridge National Laboratory (2011) and represents the great-circle distance between county centroids. To determine the great-circle distance county centroids (latitude, longitude) are first established. Then, the central angle of each centroid is determined. Finally, the great-circle distance is calculated by multiplying the Earth radius and the central angle. So, the distance is projected on the Earth sphere plane Oak Ridge National Laboratory (2011). The great-circle distance is the most commonly used distance measure in the large literature on the gravity model of international trade (Disdier and Head 2008). Even though great-circle distance is a simplification of transport pathways, the gravity model of international trade does not perform better when the actual geography of transportation is used (Disdier and Head 2008).

The likelihood and mass flux of food transfers originating from a county is related to its production of these goods. County level production ($) for unprocessed agricultural commodities (SCTG 1-4) come from US Department of Agriculture (2014). Production values of each crop or livestock category originating within a county were aggregated to their corresponding SCTG code. Similarly, the county level production of processed agricultural and food goods (SCTG 5-7) were aggregated to their respective SCTG code as described below. Production data of processed agricultural goods

Table 2. List of data sources used in this study.

Name	References	Data description	Purpose
Commodity Flow Survey (CFS) Public Use Microdata	US Census Bureau (2015b)	Survey of business shipments within the United States. FAF is largely based off this dataset, though the scope of the CFS Microdata is not as broad as that of the FAF dataset. However, the CFS Microdata contains greater shipment detail, including the NAICS industry responsible for the shipment.	This dataset allowed for pairing of commodity transfers to specific industries.
Freight Analysis Framework (FAF) Version 4	Oak Ridge National Laboratory (2015)	Data detailing freight movement between 132 major metropolitan areas and remainder of states (i.e. FAF Zones), as well as eight international import/export regions.	FAF commodity transfers are used to constrain county transfers. The sum of county transfers must equal that of the FAF Zone that they belong.
US Census Bureau 2012 Economic Census	US Census Bureau (2015a)	Provides county level economic data by industry, including employment and the value of industry output.	The Economic Census was used to determine production of processed agricultural goods and the total production output of all industries using agricultural goods as production inputs. These data were used in our gamma mixture hurdle model for link prediction and assigning flow strength.
US Department of Agriculture 2012 Census of Agriculture	US Department of Agriculture (2014)	Agricultural production data for each crop or livestock type at the county scale.	The Census of Agriculture was used to determine county level production values for each crop and livestock. These data were used in our gamma mixture hurdle model for link prediction and assigning flow strength.
Input–Output (I–O) Accounts Data	US Bureau of Economic Analysis (2014)	These data detail supply chain input requirements for each industry per unit of their output.	Direct requirement coefficients from the I-O accounts were multiplied by production data to determine the commodity input requirements of each industry, as well as end consumers. A county's total input requirement of a commodity across all industries and end consumers represents its total consumption of that good. This is used in our gamma mixture hurdle model for link prediction and assigning flow strength.
County-to-County Distance Matrix and Network Impedance	Oak Ridge National Laboratory (2011)	Matrix of distances and impedances between county centroids via different transportation methods.	The linear programming algorithm used this matrix to minimize transportation cost.
Personal Income	US Bureau of Economic Analysis (2017)	Personal income data per county.	When paired with the input–output data tables, this was used to help determine final consumer demand of different commodities within a county.
Port Trade	US Census Bureau (2018)	Value ($) and mass (kg) trade data for international ports of the United States.	Trade data to/from these ports was used to better capture transit hubs in the gamma mixture model.

originating from a specific NAICS food processing industry comes from US Census Bureau (2015a).

Other statistical information is required to determine the destination of food flows. The 2012 CFS Public Use Microdata (US Census Bureau 2015b) and the United States Bureau of Economic Analysis input–output accounts data (US Bureau of Economic Analysis 2014) were used to statistically determine the production and attraction of food within our machine learning algorithm (see below). The CFS Microdata utilizes the same survey data as the CFS dataset but provides greater shipment detail than the standard CFS data. Importantly, one additional detail included in the CFS Microdata is the North American Industry Classification System (NAICS) code of the industry producing and shipping the good. This additional information enables us to relate the SCTG code of a transported commodity to the NAICS industry producing the commodity. Since the CFS Microdata does not provide a NAICS code for raw agricultural and food goods (SCTG 01-04),

we manually matched the production of individual crops or livestock reported by US Department of Agriculture (2014) to the SCTG code to which it belongs (a listing of goods within each SCTG can be found at https://census.gov/econ/cfs/2012/2012_manual.pdf). The SCTG-NAICS crosswalk table we created (provided in the online supporting information (SI) available at) stacks.iop.org/ERL/14/084011/mmedia was paired with input–output accounts data to determine an industry's use of each SCTG as input in its production process. Input–output tables show to what degree the production (output) of one industry is used as input to another industry. Using the crosswalk table we created, we aggregate industry output within the table to its corresponding SCTG code to match the FAF4 data set. This procedure allows us to restrict data used within our machine learning algorithm to variables that have been established as relevant to the production or consumption of each SCTG good. This ensures that our model maintains realism.

Some agricultural (US Department of Agriculture 2014) and business production data (US Census Bureau 2015a) are suppressed by the data collecting agency if their release may reveal information on an individual producer. Suppressed data records are not removed from the data set, but instead flagged, indicating there are limited producers within that geographical area. Data suppression is more prevalent at the county spatial scale and among specialty producers. For example, artichoke (a specialty crop) production in Linn County, Oregon is flagged since reporting this data would reveal information specific to the only artichoke farmer in the county. When suppressed values arise in the data sets, the geographical and industry/product hierarchical structure of the data is exploited to estimate these suppressed values. The artichoke production of the sole farmer in Linn County, for example, was estimated by subtracting the sum of all artichoke production in Oregon counties from the state-level production value provided by US Department of Agriculture (2014). The difference between the state total and the sum of all counties is uniformly distributed amongst all Oregon counties with suppressed artichoke production records.

Industrial production records have other data fields that can help us further refine our estimates of suppressed production values. US Census Bureau (2015a) provides employment records for each industry within a county, which can be used to help estimate suppressed production output. Employment data is not used directly within our model. Instead, it is used to estimate the production output (which is used within our model as an input) when this production data has been suppressed. Employment data is more widely reported and is not subject to as strict data suppression requirements as production data. For each industry, we exploited the numerous instances when both production and employment data existed to establish coefficients relating the number of employees working within an industry to the production of that industry. These industry-specific coefficients were applied to employment records to estimate production when production data was suppressed within a given county. Relationships between production output and employment were established for every industry based on the large number of records where both values were provided. This allowed us to estimate production for counties with limited industrial activity. While similar approaches have been applied in the literature (Isserman and Westervelt 2006, Smith *et al* 2017, Marston *et al* 2018), our study would nonetheless benefit from a complete data set.

Port trade data is retrieved from the Census Bureau USA Trade database (https://usatrade.census.gov/index.php) for the year 2012. The values ($) and mass (kg) for both sea and air ports are provided (US Census Bureau 2018). Value flows were ultimately used due to significantly more data availability as compared to mass. While land ports are not specifically mentioned, many of the reported ports are US Customs and Border Patrol crossings on US land borders (such as along the Northern borders of North Dakota and Montana), implying that land ports are included in the database. Commodities in the port trade database are reported using the HS coding system. For consistency with FAF flow data, a crosswalk was created to convert from HS to SCTG codes. The Python geocoder library (http://geocoder.readthedocs.io/) was then used to determine latitude and longitude coordinates for each port. Some ports, such as 'Low Value (Port)', did not have locations and were consequently removed. A spatial join was finally used to determine which county each port is in, resulting in 331 ports in 228 counties contributing inflows and outflows of SCTGs 1 through 7 in the US. We use this port data in our algorithm (see section 2.2) to boost fluxes to/from transit hubs that do not directly correspond to production/consumption flows.

2.2. The Food Flow Model

We develop the Food Flow Model to estimate food flows between counties in the United States. Our goal is to estimate F, which is a weighted and directed matrix of food flows between all county pairs (i.e. for all 9869 022 potential links within the nation). F provides flows from an origin (o) to a destination (d) county.

Figure 3 provides a schematic of our algorithm. Our approach is based on supervised-learning and linear programming methods with mass balance and commodity network properties as constraints. Our algorithm is comprised of three main steps. Step 1: train a gamma mixture hurdle model for food commodity flows between FAF zones. Refer to the SI for the list of variables used in our training algorithm. Step 2: simulate the commodity flow potentials between counties using the model obtained in Step 1. Step 3: use linear programming to minimize the distance of food flows between counties. The major assumption employed in our modeling framework is that the supervised learning

model established for flows between FAF zones is representative of flows between counties.

2.2.1. Step 1: develop gamma mixture hurdle model at FAF scale

In step 1, we train a gamma mixture hurdle model for commodity flows between FAF zones. A hurdle model is a two-part model in which one process is specified for zeros and another process for positive values. The idea is that positive values occur once a hurdle is cleared. Our hurdle model uses logistic regression to predict the presence or absence of a link (i.e. to determine the binary adjacency matrix (A) corresponding to F) and a gamma mixture model to estimate the mass of existing links (i.e. to assign values to the '1s' in A and obtain the weighted F matrix). So, our gamma mixture hurdle model is composed of two major components: (i) a logistic regression model for link topology (i.e. the A matrix) and (ii) a gamma mixture model for flow strength (i.e. the F matrix).

(i) Logistic regression model: food flow networks exhibit connectivity distributions that follow the generalized exponential-binomial distribution across scales (Konar *et al* 2018). This indicates that link generation can be modeled as a two-step process. First, the probability (p) is sampled from a prior generalized exponential distribution. Second, a coin is flipped to obtain a value. If the flipped value is greater than p then a connection is made between an origin and destination node. However, the generalized exponential prior distribution does not provide enough restriction to estimate the presence or absence of all links in our system. So, logistic regression can be used to take additional geographic and economic features of the counties into account.

We use the binary logistic regression model to estimate the probability of a binary response based on available predictor (or independent) variables (Cox 1958, Walker and Duncan 1967). In binary logistic regression, the outcome is coded as '0' or '1'. Here, an outcome of '0' indicates that no link is present between two nodes, while an outcome of '1' indicates that a link exists. We use supervised learning to determine the functional form of the logistic model for each SCTG food group. Supervised learning is a machine learning task to learn the function that maps an input to an output. This learning process infers a function from training data (Mohri *et al* 2012). Here, we use the available FAF zone food flow data as our training data set. We use link-level FAF food flow data, so 17 292 data points are available in our training data set. Using supervised learning, a logistic regression model is established for each SCTG food commodity group. The logistic regression model for each SCTG group is provided in the SI. So, the logistic regression model determines the presence or absence of a food flow link. If a food flow link is predicted to exist, then the 'hurdle' has been cleared. Once the hurdle has been cleared, then we use a gamma regression model (part (ii)) to estimate the mass transfer on that link. In this way, the logistic regression model determines if the 'hurdle' has been passed and the gamma regression model will only proceed after getting a positive result from the logistic regression.

The area under the curve (AUC) metric with ten-fold cross-validation is used to evaluate model performance. AUC measures the entire two-dimensional area below the receiver operating characteristic (ROC) curve from (0, 0) to (1, 1). AUC provides an aggregate measure of performance across all possible classification thresholds. AUC is desirable because it is scale-invariant and classification-threshold-invariant. It measures how well predictions are ranked, rather than their absolute values. Additionally, AUC measures the quality of the model's predictions irrespective of what classification threshold is chosen. AUC values range from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. A score of 0.5 is no better than random chance. There is a tradeoff between precision and overfitting. A score of 0.9 indicates a very good model, but a score of 0.9999 may be too good to be true and will indicate overfitting. AUC metrics are provided in the SI and are within the range of 0.78–0.93 across commodity categories.

(ii) Gamma mixture model: food mass flux distributions follow the gamma distribution across scales (Konar *et al* 2018). The gamma distribution is generated from the homogeneous Poisson process with a constant rate of success (Boland 2007). For this reason, the commodity flow process can be modeled using a Poisson process. Conceptually, this implies that food commodities will be transported from the origin to the destination until k (shape of gamma distribution) effective units of the food commodity are delivered. To understand this 'effectiveness', we can consider the example of animal feed. A feed manufacturer needs to produce a certain amount of feed containing k units of corn. The origin ships corn, but not all corn ends up in the feed. Some corn might be lost during transport, some corn might be sent to other manufacturing plants or used for other purposes besides feed, and some corn might be re-exported. So, the corn that finally reaches the feed manufacturer is only a fraction, and this fraction is the success rate. Here, we approximate this success rate as a constant within each food commodity category. The gamma regression model for each SCTG group is provided in the SI.

For most food commodity groups about 5% of the flows exceed the upper bound of the 95% confidence interval of our gamma regression model. These outliers correspond to major transportation hubs within the US (e.g. ports). These outliers lead us to consider transit hubs as an additional attribute for some key nodes. We use the port trade data from the Census Bureau USA Trade database for these hubs (see section 2.1 and table 2). The same process is employed to generate the second gamma model. In this way, we develop a gamma mixture model (Llordén 2017). In our gamma mixture model, there are two gamma

regression models with different feature spaces. So, our gamma mixture hurdle model combines (i) supervised learning to established a logistic regression model for link prediction and (ii) a gamma mixture model to estimate the mass of estimated links, taking transit hubs into account.

2.2.2. Step 2: simulate commodity flow potentials between counties

In step 2, we simulate the commodity flow potentials between counties given the model developed for FAF flows in step 1. The logistic regression model developed for each SCTG food commodity is used to decide the topology of food flows at county level. If the probability is greater than the selected threshold, a link is assigned. Next, the mass flux of food flows is estimated for existing links. Gamma regression (without total importing information of international ports) is used to estimate the expected value of the food flows between counties. We generate flow potentials as random variables sampled from a gamma distribution with the expected values of these food flows as the scale of the gamma model. If there exists potentials between FAF zones summing to a value smaller than reported flow between these FAF zones, the gamma regression with total importing information of international ports is used to re-estimate the expected values between these counties. If the total of flows among counties between FAF zones are still smaller than the reported value, scaling is used.

2.2.3. Step 3: solve system with linear programming

In step 3, we apply a linear programming process to solve the system. Our linear program takes the flows between FAF zones as mass balance constrains and the potentials estimated in step 2 as inequality constraints. Then, the linear program minimizes the transportation distance of food flows between counties (Klein 1967, Ahuja *et al* 1993). Given the network topology and flow potentials estimated in steps 1 and 2, we estimate the flows between counties with minimization of distance as the objective function and flow values reported at FAF level as equality constraints. In step 2, we guaranteed that the sum of the flow potentials among counties within an FAF zone is always greater than or equal to the corresponding FAF zone flows. So, there is always a solution to this linear programming problem.

The solution to our linear programming system minimizes transport distance while ensuring mass balance between counties modeled within an FAF zone. In this way, our approach builds on the gravity model of trade in which distance (which typically correlates with costs) is inversely related to trade flows (Disdier and Head 2008). Note that our model framework is not as strongly influenced by the distance minimization assumption as many other studies (e.g. Smith *et al* 2017) that rely on this assumption since our model is additionally bounded by the FAF data.

Note that the Food Flow Model estimates self loops at the county scale. This is because self loops exist in the FAF data. For example, the remainder of California reports a flow to the remainder of California, such that the remainder of California is both the origin and destination of the flow.

2.3. Global sensitivity and uncertainty analysis

A global sensitivity and uncertainty analysis (GSUA) can help to determine the variables that are most influential in model output (Saltelli *et al* 2004, Ludtke *et al* 2007, Convertino *et al* 2014, Servadio and Convertino 2018). Here, we implement the Fourier amplitude sensitivity test (FAST) method to calculate the contribution of each input variable to the output variance (Cukier *et al* 1973, Saltelli *et al* 1999). The main advantage of the FAST method is that it is robust for relatively small sample sizes (Cukier *et al* 1973). Additionally, the FAST method is computationally efficient (Saltelli *et al* 1999). GSUA methods may require a pre-screening method, such as the Morris method, to reduce the number of variables (Convertino *et al* 2014, Servadio and Convertino 2018). However, here, we have a relatively small number of variables, so we do not require a pre-screening and are able to directly perform GSUA with FAST.

The 'first-order index', as named in Sobol's method, measures the contribution of input X alone to the output variance. In this metric, no impact through interaction is considered. For example, in a system $y = 3x_1 + x_1x_2$, the first-order sensitivity of x_1 only considers the effect of $3x_1$ and $x_1E(x_2)$, where x_2 has been averaged out. In comparison, the 'total-order index' of x_1, measures the contribution of input variance to the output variance, including all variance caused by its interactions, of any order, with any other input variables (Homma and Saltelli 1996). So, the effect of all the group of variables that contain x_1, would consider impact of both $3x_1$ and x_1x_2.

First, we fit a lognormal probability distribution function (PDF) to each input variable. There are two main reasons for fitting the input variables with the lognormal distribution: (1) all input variables are highly right skewed. It is common practice to log transform right skewed data before regression. (2) In the trade economics literature, the gravity model of trade is a prevalent empirical model to describe trade systems. The gravity model of trade log transforms the input variables. The functional form of the Food Flow Model regression models are based upon the gravity model and so using the lognormal distribution enables our approach to be compared with the gravity model literature.

The lognormal PDF fit is provided in the SI. Note that the mean of the lognormal distribution is sometime negative, despite the fact that input variables to lognormal are always $\in [0, \infty)$. This is because when the input variable is <1, then log of x will be negative.

This indicates that after the input data go through log transformation, the resulting mean is negative, or that there are many values less than 1. In some instances, this may be due to counties that do not have data, where we replace zero with 0.1^{-100}, so that it can be log transformed. Notice that sometimes the same variable for different SCTGs have different PDFs because they represent different commodities (e.g. vegetable versus grains).

For GSUA, the input random variables are assumed to follow the lognormal distribution. The minimum and maximum value of each random variable are taken as bounds. The simulation iterations are increased by 1000 until the total-order index's change between the previous and current session is within 5% for any variable factor.

3. Results and discussion

There are 132 nodes in the FAF census data and 3132 nodes in the county model results (see table 3). So, we do not model 10 of the 3142 counties in the United States due to limited data for these counties (refer to the SI for a list of these counties). There are 11 678 links in the FAF data out of a potential 17 292 links, leading to a density of 0.675. The Food Flow Model estimates 161 394 non-zero links at the county scale out of a potential 9869 022 links. This means that the density of the county scale food flows is 0.016. The inferred network density at the county scale is much less than the empirical density at the FAF scale. However, it makes sense that the density at the county scale is lower than the FAF scale, since this finer spatial resolution makes it unlikely that most counties would connect with one another directly and would instead transit through hubs. Of importance, note that the mass balances for each SCTG commodity class between the county and FAF spatial scales as required (see table 3).

Table 3. Network properties of food flows within the United States. Properties are provided for each SCTG group at the FAF and county spatial resolution. The mass flux of FAF scale self loops are included, as this mass is distributed amongst counties within those FAF zones.

FAF SCTG	# Nodes	# Links	Mass (kg)	Density
1	132	1454	89.35E+9	0.085
2	132	1441	789.16E+9	0.083
3	132	4535	397.51E+9	0.262
4	132	3491	258.60E+9	0.201
5	132	4681	76.76E+9	0.271
6	132	4762	101.42E+9	0.275
7	132	8924	559.00E+9	0.516
Total	132	11 678	2.27E+12	0.675
County				
1	2945	14 474	89.35E+9	0.002
2	2904	19 563	789.16E+9	0.002
3	2491	16 272	397.51E+9	0.003
4	2594	18 799	258.60E+9	0.003
5	2817	30 670	76.76E+9	0.004
6	2628	30 842	101.42E+9	0.004
7	3050	71 826	559.00E+9	0.008
Total	3132	161 394	2.27E+12	0.016

3.1. County-scale food flows

Figure 4 maps food inflows and outflows at the FAF and county spatial scales. The spatial trends compare well between FAF and county spatial scales. For example, note that California and the Great Lakes region are major outflow locations in the FAF data (see figure 4(B)). Counties within these FAF areas are also locations of high food outflow in the nation (see figure 4(D)). Similarly, the counties that are estimated to receive the most inflows of food correspond to the locations of FAF zones with high food receipts (compare figure 4(A) with figure 4(C)). This indicates that the Food Flow Model is maintaining the broad spatial trends observed at the FAF spatial scale as designed. Note that the mass transfer at each scale is different, as indicated by different scales on the color bars. The masses being transferred at the county scale are smaller than at the FAF level, because the mass at the FAF level has been distributed to counties, and we do not expect a single county within an FAF zone to transfer the entirety of the food mass. However, we are now able to infer how food flows are distributed across counties, which we are not able to observe at the FAF scale.

Table 4 ranks the top outflow and inflow locations by spatial scale. Our model estimates that several California counties are the largest in terms of outflows and inflows. For example, Los Angeles county is predicted to be the largest origin and destination node at the county scale, despite the fact that the remainder of Iowa FAF zone is the largest origin and destination node at the FAF scale (refer to table 4). This indicates that the large remainder of Iowa link is more evenly distributed amongst the counties within Iowa, while the mass flux within the state of California is distributed in a fairly heterogeneous manner amongst its counties. This is because of the high heterogeneity in production and consumption within California. This is also a function of the linear programming algorithm that minimizes distance. Distances between counties in northern and southern California are larger than distances across the state of Iowa, for example. Our model estimates more local flows in California since the linear program objective function enacts a heavy penalty for transporting food large distances within the state. Additionally, counties in the western portion of the United States, including California, are larger than counties in the east, which also leads to more aggregation at the county spatial scale.

Figure 5 maps food flows at the FAF and county spatial scales. Links are shown for all FAF flows and for the largest 5% of county estimates. These maps depict

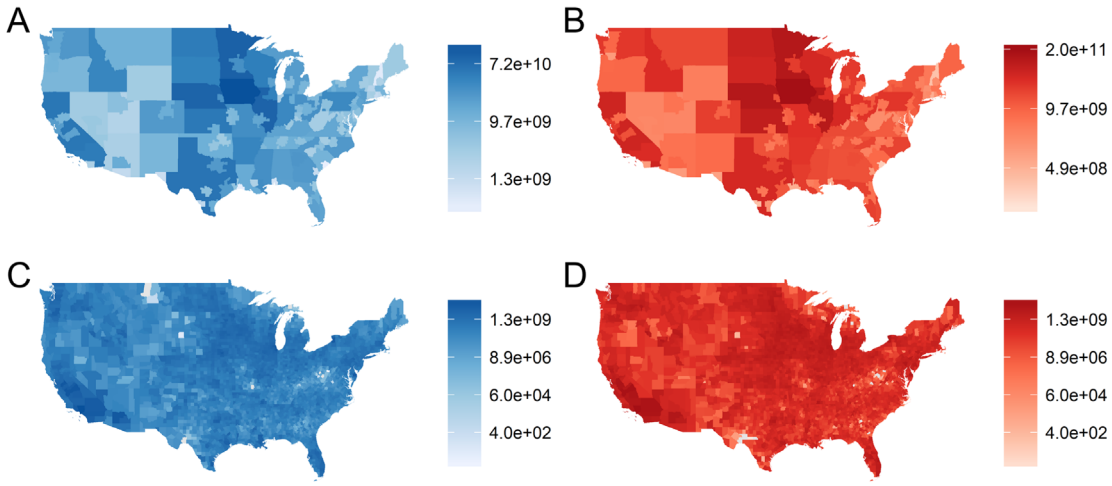

Figure 4. Maps of total food flows (tons) within the United States. Maps show (A) FAF inflows, (B) FAF outflows, (C) county inflows, and (D) county outflows. Maps for specific food commodity groups are provided in the SI.

Table 4. Ranking of total food flows within the United States by mass (kg). The top 10 food outflow and inflow FAF zones and counties are provided. Note that self-loops are included. The ranking for specific food commodity groups is provided in the SI.

FAF				
Rank	Outflow	Mass (kg)	Inflow	Mass (kg)
1	Remainder of Iowa	1.97E+11	Remainder of Iowa	1.60E+11
2	Remainder of Nebraska	1.27E+11	Remainder of Minnesota	1.06E+11
3	Remainder of Minnesota	1.26E+11	Remainder of Illinois	1.00E+11
4	Remainder of Illinois	1.17E+11	Remainder of Nebraska	9.91E+10
5	Remainder of Kansas	9.13E+10	Remainder of Texas	6.66E+10
6	Remainder of North Dakota	6.98E+10	Remainder of California	6.23E+10
7	Remainder of California	6.32E+10	Remainder of Kansas	5.81E+10
8	Remainder of South Dakota	5.86E+10	Los Angeles-Long Beach, CA CFS Area	5.48E+10
9	Remainder of Wisconsin	5.32E+10	Remainder of North Dakota	5.48E+10
10	Remainder of Texas	5.23E+10	Remainder of Wisconsin	5.09E+10
County				
1	Los Angeles County, CA	1.66E+10	Los Angeles County, CA	2.19E+10
2	Fresno County, CA	1.24E+10	Fresno County, CA	1.23E+10
3	Stanislaus County, CA	9.92E+09	Stanislaus County, CA	1.18E+10
4	San Bernardino County, CA	9.78E+09	Maricopa County, AZ	1.07E+10
5	San Joaquin County, CA	8.88E+09	Orange County, CA	9.54E+09
6	Merced County, CA	8.86E+09	Riverside County, CA	8.71E+09
7	Riverside County, CA	8.69E+09	Erie County, NY	8.53E+09
8	Tulare County, CA	7.97E+09	Cook County, IL	8.47E+09
9	Kern County, CA	5.77E+09	Douglas County, NE	8.27E+09
10	Maricopa County, AZ	5.72E+09	Sussex County, DE	7.78E+09

aggregate food flows and the general spatial trends between the FAF and county spatial scales compare well. For example, note that the strong connectivity between the corn/soy belt and the port of New Orleans exists in both the FAF data (see figure 5(A)) and the county modeled results (see figure 5(B)). Similarly, the links between the New York area and the Great Lakes, as well as the connections from the grain belt to California, are shown in both figures 5(A) and (B). The density is much higher for the FAF data than inferred county results (refer to table 3). However, this is sensible, since the spatial scale is so much larger (by definition) in FAF, there will be more connectivity. The county flow results were additionally pruned to exclude links with fluxes <1 kg, further reducing the estimated density at this scale.

Note the prevalence of self-loops in both the FAF and county results (see table 5). For example, the transfer of food from Los Angeles County, CA to Los Angeles County, CA is one of the largest links at the county scale. Note that the Food Flow Model estimates a flow of food, which occurs each time a commodity is transformed (i.e. corn into corn meal into biscuits). Cities are important food manufacturers who process food items

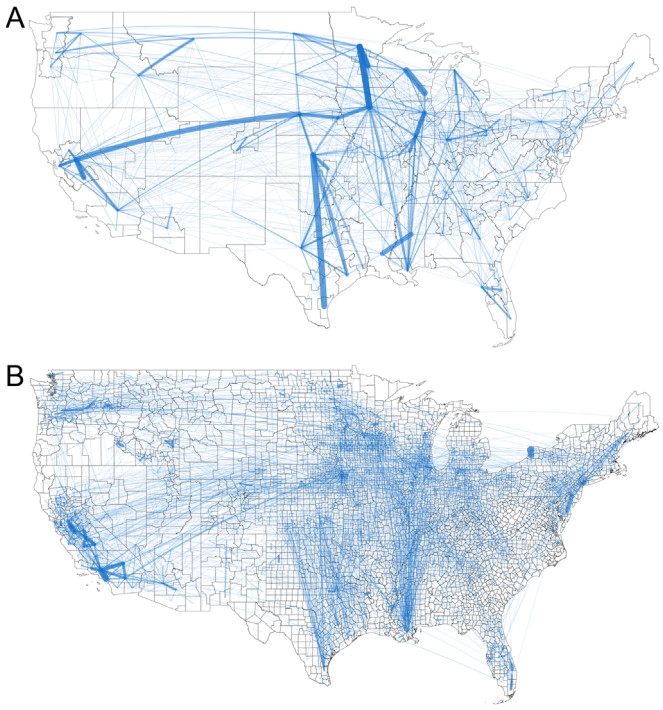

Figure 5. Maps of food flow networks within the United States. Maps depict total food flows (tons) for the (A) FAF and (B) county scale. Links are shown for all FAF data and for the largest 5% of county links.

Table 5. Ranking of total food flows within the United States by mass (kg). The top 10 links at the FAF and county scales are provided. Note that self-loops are included. The ranking for specific food commodity group is provided in the SI.

FAF		
Rank	Link	Mass (kg)
1	Remainder of Iowa → Remainder of Iowa	1.32E+11
2	Remainder of Nebraska → Remainder of Nebraska	8.43E+10
3	Remainder of Minnesota → Remainder of Minnesota	7.71E+10
4	Remainder of Illinois → Remainder of Illinois	7.10E+10
5	Remainder of North Dakota → Remainder of North Dakota	4.72E+10
6	Remainder of Kansas → Remainder of Kansas	4.31E+10
7	Remainder of South Dakota → Remainder of South Dakota	4.13E+10
8	Remainder of Texas → Remainder of Texas	3.80E+10
9	Remainder of Idaho → Remainder of Idaho	3.13E+10
10	Los Angeles-Long Beach, CA CFS Area → Los Angeles-Long Beach, CA CFS Area	3.09E+10
County		
1	Broomfield County, CO → Broomfield County, CO	5.21E+09
2	Los Angeles County, CA → Los Angeles County, CA	4.23E+09
3	Los Angeles County, CA → Orange County, CA	3.93E+09
4	Niagara County, NY → Erie County, NY	3.92E+09
5	Merced County, CA → Stanislaus County, CA	3.49E+09
6	San Diego County, CA → San Diego County, CA	3.25E+09
7	Erie County, NY → Erie County, NY	2.78E+09
8	Camden County, NJ → Sussex County, DE	2.72E+09
9	San Joaquin County, CA → Stanislaus County, CA	2.46E+09
10	Stanislaus County, CA → Merced, CA	2.35E+09

from one form to another. This is especially true of Los Angeles County, whose food manufacturing industry produced nearly $16 billion in goods in 2012, the largest of any county in the United States (US Census Bureau 2015a). Further, Los Angeles brings in food from other countries and agricultural production hubs around the United States. In fact, the Los Angeles FAF zone is the second largest food importer behind only the New Orleans FAF zone. Together, the large food manufacturing presence and sizeable international

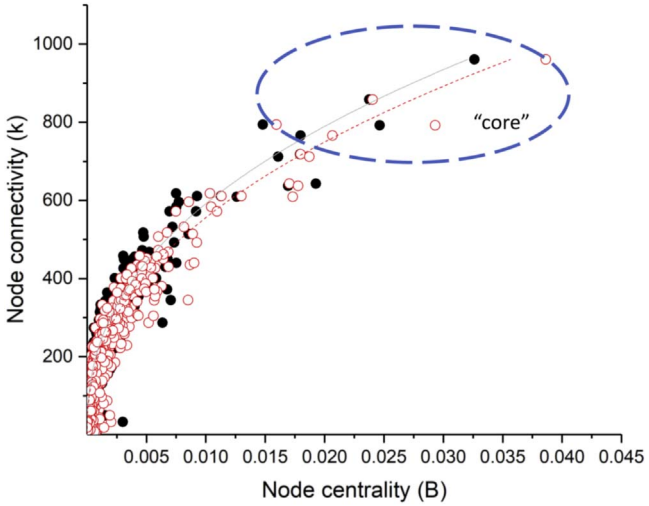

Figure 6. Relationship between node connectivity and centrality. Node connectivity is measured by node degree, k. Node centrality is measured by between-ness centrality, B. Red points indicate directed food flows, while black points indicate undirected food flows. The relationship is described by $B = a \times k^b$ as in other empirical studies at different spatial scales. A 'core' group of counties is evident: San Berndardino, CA; Riverside, CA; Los Angeles, CA; Shelby, TN; San Joaquin, CA; Maricopa, AZ; San Diego, CA; Harris, TX; and Fresno, CA.

imports explain why Los Angeles County is the largest self-loop at the county scale.

Over half of the 10 largest links are estimated to be within California (see table 5). This model result is sensible due to the large mass fluxes reported in the FAF data combined with the large spatial heterogeneity in production and attraction factors for food in California. Even though other FAF zones tend to transfer larger masses than those within California, these flows are distributed across many more smaller counties. California counties can be 1–2 orders of magnitude larger than counties in the eastern United States and exhibit greater heterogeneity in production and consumption than Midwestern counties. For example, the largest FAF transfer is the self-loop for remainder of Iowa (to itself). Due to relatively limited diversity in crop type and production patterns across Iowa, it is unsurprising that county level food transfers within Iowa are relatively evenly spread. The homogeneous production and distribution would inhibit a handful of counties from transferring all food within the state. In this way, the more heterogenous distribution of production and consumption within California leads to more concentrated food flows.

The Food Flow Model preserves known patterns at other spatial scales. Figure 6 shows the relationship between node connectivity (i.e. degree, k) and centrality (i.e. between-ness centrality, B). The relationship between B and k has been presented as $B = a \times k^b$ in empirical studies at different spatial scales (Ercsey-Ravasz *et al* 2012, Lin *et al* 2014, Konar *et al* 2018). This relationship is preserved in our estimated county flow network with adjusted $R^2 = 0.93$ ($a = 2.2 \times 10^{-9}$; $b = 2.4$) for the undirected network and adjusted $R^2 = 0.94$ ($a = 4.1 \times 10^{-9}$; $b = 2.3$) for the directed network. A 'core' group of counties is evident which have both high connectivity and centrality and are critical to the structure of the domestic food flow network. These core counties are: San Berndardino, CA; Riverside, CA; Los Angeles, CA; Shelby, TN; San Joaquin, CA; Maricopa, AZ; San Diego, CA; Harris, TX; and Fresno, CA. Importantly, the Food Flow Model does not prescribe this relationship; it naturally emerges from our algorithm. Figure 6 demonstrates that the Food Flow Model is able to generate known macroscopic properties of other food flow networks (e.g. as presented by Konar *et al* 2018).

Table 6. Metrics to compare model performance of the Food Flow Model with Smith *et al* (2017). Note that maize fluxes for the Corn Belt only are considered. The simple matching coefficient (SMC), R-squared (R2), mean absolute error (MAE), and root mean squared error (RMSE) metrics are provided for inter-county links and county-scale inflows and outflows.

	SMC	R2	MAE	RMSE
Links	1	0.07	126, 178, 325	202, 361, 790
Inflow	1	0.04	405, 160, 210	734, 981, 931
Outflow	1	0.46	310, 642, 341	469, 188, 425

3.2. Comparison with literature

We compare our results with the county-scale corn flows modeled by Smith *et al* (2017). To our knowledge, this is the only other information on county-scale food flows in the United States. Smith *et al* (2017) use a transportation optimization model to estimate corn flows between US counties. To compare our results, we transform our estimates of SCTG 02 to estimates of corn by multiplying the SCTG 02 flows by the fraction of corn grains produced in each origin county as compared to total grain production. Grains here include the following crops: barley, buckwheat, corn (grain), corn (silage), millet (proso), oats, rice,

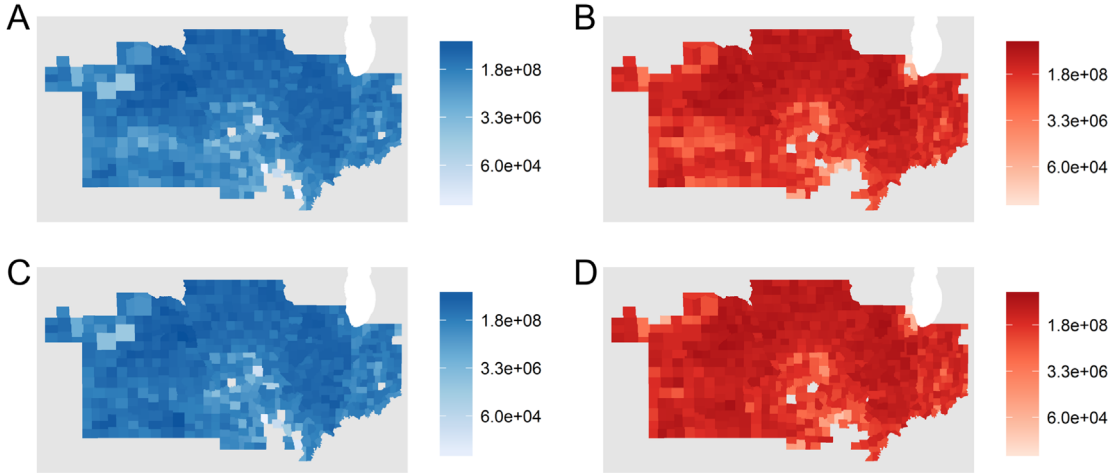

Figure 7. Comparison of maize flow [tons] maps for the Corn Belt of the United States. Maps show (A) maize inflows for our model, (B) maize outflows for our model, (C) maize inflows from Smith *et al* (2017), and (D) maize outflows from Smith *et al* (2017).

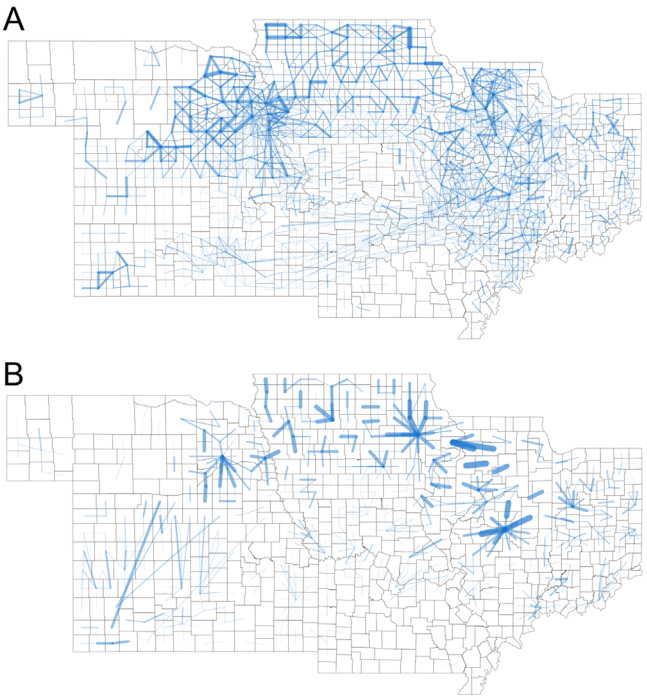

Figure 8. Comparison of maize flow networks [tons] for the Corn Belt of the United States. (A) The Food Flow Model for SCTG 02 scaled by maize production within each county. (B) Maize flows from Smith *et al* (2017). Links represent maize flows between counties.

rye, sorghum (grain), sorghum (silage), triticale, wheat, and wild rice. We compare results for Corn Belt states (Illinois, Indiana, Iowa, Kansas, Missouri, and Nebraska) to avoid inaccuracies in states which produce large quantities of other grains.

Table 6 presents metrics to compare model output between the Food Flow Model and Smith *et al* (2017). The simple matching coefficient (SMC) (Gower 1971) indicates that the models identically estimate the presence or absence of links within the Corn Belt (SMC = 1). Both models estimate the existence of links between most county pairs in the Corn Belt, albeit many with very small values. The root mean squared error, mean absolute error, and R-squared (R2) metrics are also presented in table 6. These metrics additionally consider the intensity of the fluxes. For this reason, these metrics are more strict in assessing model performance than SMC. Note that the R2 value is highest for the county-scale outflows. This indicates that the two models have the most agreement on outflows, likely due to the quality of the production statistics that drives estimation of this variable.

Figure 7 maps total inflows and outflows for each Corn Belt county. Note that both inflow and outflow maps share a common scale. The spatial trends compare remarkably well between the two models. Figure 8 shows how our corn flows compare to the

Table 7. Comparison of maize flows for the Corn Belt in the United States. Note that values for this study (without corn production scaling) are equivalent to the data reported by the Freight Analysis Framework (FAF).

This study (with corn production scaling)				
State	Outflows	Inflows	Intraflows	All Flows
IL	2.24E+09	4.23E+09	6.54E+10	7.19E+10
IN	2.20E+09	5.49E+08	2.24E+10	2.52E+10
IA	6.75E+09	2.31E+09	7.29E+10	8.19E+10
KS	4.40E+09	3.71E+09	1.98E+10	2.79E+10
MO	1.43E+09	2.94E+09	9.48E+09	1.39E+10
NE	5.30E+09	8.58E+09	6.48E+10	7.87E+10
This study (without corn production scaling)				
IL	2.57E+09	5.17E+09	7.09E+10	7.87E+10
IN	2.28E+09	6.25E+08	2.32E+10	2.61E+10
IA	6.76E+09	2.32E+09	7.31E+10	8.21E+10
KS	6.53E+09	4.12E+09	4.47E+10	5.54E+10
MO	1.68E+09	3.56E+09	1.30E+10	1.83E+10
NE	5.64E+09	9.66E+09	6.73E+10	8.26E+10
Smith *et al* (2017) model				
IL	6.97E+09	6.89E+07	2.31E+10	3.01E+10
IN	6.07E+07	1.29E+09	1.46E+10	1.59E+10
IA	7.90E+08	5.61E+09	4.84E+10	5.48E+10
KS	4.04E+08	1.92E+09	8.03E+09	1.04E+10
MO	3.29E+08	8.92E+08	3.99E+09	5.22E+09
NE	1.95E+09	7.12E+08	2.46E+10	2.73E+10
FAF4				
IL	2.57E+09	5.17E+09	7.09E+10	7.87E+10
IN	2.28E+09	6.25E+08	2.32E+10	2.61E+10
IA	6.76E+09	2.32E+09	7.31E+10	8.21E+10
KS	6.53E+09	4.12E+09	4.47E+10	5.54E+10
MO	1.68E+09	3.56E+09	1.30E+10	1.83E+10
NE	5.64E+09	9.66E+09	6.73E+10	8.26E+10

flows estimated by Smith *et al* (2017). Note that maps in figure 8 share the same scale and indicate that our model has more links and larger outflows for many counties. The SI compares top 10 rankings for inflows, outflows, and links between our model and Smith *et al* (2017). Our model has more links with smaller values. In particular, our model shows more links in Nebraska than does Smith *et al* (2017), which is more concentrated in Illinois and Iowa.

Table 7 provides Corn Belt state outflows, inflows, and intra-flows (flows from a state to itself). This information is provided for FAF data, Smith *et al* (2017), our model without corn normalization, and our model with corn normalization. The total flows estimated differ between our model of corn and the Smith *et al* (2017) model. It is important to note that our model without corn normalization replicates the raw FAF data (as designed). So, our model characterizes similar spatial trends as Smith *et al* (2017), with the additional advantage of being constrained by FAF data. Importantly, Smith *et al* (2017) use their estimated corn flows to perform a spatially explicit environmental impact analysis of the US corn supply chain. This application of lifecycle assessment to spatially refined corn flows highlights potential applications of similar methods to the spatially refined estimates of all food flows provided in this paper.

Comparing our model with Smith *et al* (2017) is not a true validation exercise. Yet, it is useful to compare outcomes across these two model frameworks. Smith *et al* (2017) do not validate their model output with real-world data (since none is available), so we are unable to confirm which model most accurately represents reality. However, we believe that our food flow model improves upon the model presented in Smith *et al* (2017). This is because we constrain our results with FAF scale information. Additionally, we require our estimates to follow known properties of food flow networks at other spatial scales. We also incorporate machine learning, which is skilful in accurate estimations, albeit with limited mechanistic intuition.

3.3. Sensitivity analysis

Through the GSUA we have found that distance is the most influential variable to the output variance across all SCTGs. For example, for SCTG 03 the first order sensitivity of distance is 0.046. This means that 4.6% of

the variance of SCTG 03 flows is impacted by distance, when keeping all other variables at their average value. However, the total-order sensitivity of distance for SCTG 03 is 0.88. This indicates that 88% of the output variance is driven by the input variable distance. This is not surprising because distance has long been known to be a key factor in trade models (e.g. the gravity model of international trade) (Disdier and Head 2008). Note that the GSUA results for all SCTGs are provided in the SI.

Based on the criterion of the total-order index, all variables in our model are important. This a not surprising because we used the least absolute shrinkage and selection operator (Lasso) method to remove less important variables to avoid overfitting in our supervised learning modeling process. We also applied a significance test for each variable in our gamma regression model to remove any variables that are not significant. Some input variables have a very small first-order index. This indicates that some variables impact the result not by themselves but through interactions with other variables. Note that we choose not to include interaction terms in our model, because they would make comparison with the gravity model of trade difficult. However, future research may want to explore these interaction terms in more detail.

These GSUA findings have implications for food supply chain management. Distance plays an outsized role in the model outcome. This highlights the fact that investments in the underlying transportation system (which reduce the cost (either in financial or convenience/time terms) may enable the transmission of more food fluxes.

4. Conclusion

We developed the Food Flow Model to estimate food flows between all county pairs in the United States. To do this, we developed a novel, data-driven framework that incorporates supervised learning, mass balance, network constraints, and linear programming. This modeling framework ensures that our county-scale estimates are in accordance with available empirical information and lend additional reliability to Food Flow Model results. Our estimates of corn flows between counties in the US compare well with Smith *et al* (2017), the only other county-scale information that we are aware of. Our work contributes to the food flow modeling literature by modeling all food commodity flows, as well as incorporating the idea of network constraints and mass balance. Going forward, data collection efforts to validate agri-food supply chain models of the United States will be of increased importance.

We provide estimate of county-scale food flows for the year 2012. This was an exceptional drought year in the United States. Importantly, the drought impacts should be captured by the FAF data, as well as production and consumption data utilized within our model already, meaning our model was able to incorporate these notable conditions. However, it is possible that the regression models (determined by the supervised learning algorithm on FAF data) will be specific to each time period. Our modeling framework is general and would apply in other years; however, the Food Flow Model should be run in each new time period to ensure the most accurate results. In fact, we suggest that comparing model structure and performance in a different time period (i.e. non-drought year) is an an important area of future research.

Future work could improve the realism of our algorithm. To capture the nonlinearity between environmental variables and food flow, future models could include interaction terms and higher order polynomials into the gamma regression model. Alternatively, algorithms could utilize stacking (Wang *et al* 2011) by including the output of multiple nonlinear learners (e.g. deep learning) into a gamma regression to better capture nonlinearity while preserving the conditional gamma distribution of the output flow estimate. As another example, more realistic distance matrices could be used, such as those that are constrained by available infrastructure. For example, future research could utilize distances between counties based upon the roadway network, rather than shortest paths. In fact, future research could take advantage of the mode information provided by FAF to further resolve these food flow estimates to specific infrastructure networks (i.e. road, rail, waterway). If this is accomplished, an inter-connected network model could be developed to reveal potential vulnerabilities and resiliencies of the national food supply chain. Additionally, future research could combine these detailed food flow estimates with high-resolution footprint estimates to evaluate the water, carbon, and nutrient footprint of the national food supply chain in the United States.

Critically, we make our inferred food flows freely available and publicly accessible with this paper. In this way, we provide transparency to our work and enable future researchers to build upon our results. The Food Flow Model output is provided in the SI.

Acknowledgments

This material is based upon work supported by the National Science Foundation Grant No. ACI-1639529 ('INFEWS/T1: Mesoscale Data Fusion to Map and Model the US Food, Energy, and Water (FEW) System'), EAR-1534544 ('Hazards SEES: Understanding Cross-Scale Interactions of Trade and Food Policy to Improve Resilience to Drought Risk'), and CBET-1844773 ('CAREER: A National Strategy for a Resilient Food Supply Chain'). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not

necessarily reflect the views of the National Science Foundation. All data sources are detailed in table 2 and are publicly accessible. We gratefully acknowledge these sources, without which this work would not be possible.

ORCID iDs

Megan Konar https://orcid.org/0000-0003-0540-8438

References

Ahuja R K, Magnanti T L and Orlin J B 1993 *Network Flows: Theory, Algorithms, and Applications* (London: Pearson)

Boland P J 2007 *Statistical and Probabilistic Methods in Actuarial Science* (New York: Chapman and Hall) (https://doi.org/10.1201/9781584886969)

Convertino M, Muoz-Carpena R, Chu-Agor M L, Kiker G and Linkov I 2014 Untangling drivers of species distributions: global sensitivity and uncertainty analyses of MaxEnt *Environ. Modelling Softw.* **51** 296–309

Cox D 1958 The regression analysis of binary sequences *J. R. Stat. Soc.* B **20** 215–42

Cuéllar A D and Webber M E 2010 Wasted food, wasted energy: the embedded energy in food waste in the United States *Environ. Sci. Technol.* **44** 6464–9

Cukier R I, Fortuin C M and Shuler K E 1973 Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I theory *J. Chem. Phys.* **59** 3873–8

Dalin C and Rodriguez-Iturbe I 2016 Environmental impacts of food trade via resource use and greenhouse gas emissions *Environ. Res. Lett.* **11** 035012

Dang Q, Lin X and Konar M 2015 Agricultural virtual water flows within the United States *Water Resour. Res.* **51** 973–86

Deryugina T and Konar M 2017 Impacts of crop insurance on water withdrawals for irrigation *Adv. Water Res.* **110** 437–44

Disdier A-C and Head K 2008 The puzzling persistence of the distance effect on bilateral trade *Rev. Econ. Stat.* **90** 37–48

Ercsey-Ravasz M, Toroczkai Z, Lakner Z and Baranyi J 2012 Complexity of the international agro-food trade network and its impact on food safety *PLoS One* **7** e37,810

FAO 2013 Staple foods: what do people eat? (http://fao.org/docrep/u8480e/U8480E07.htm#Staple%20foods%20What%20do%20people%20eat)

Gower J 1971 A general coefficient of similarity and some of its properties *Biometrics* **27** 857–71

Homma T and Saltelli A 1996 Importance measures in global sensitivity analysis of nonlinear models *Reliab. Eng. Syst. Saf.* **52** 1–17

Isserman A M and Westervelt J 2006 1.5 million missing numbers: overcoming employment suppression in county business patterns data *Int. Reg. Sci. Rev.* (https://doi.org/10.1177/0160017606290359)

Klein M 1967 A primal method for minimal cost flows with applications to the assignment and transportation problems *Manage. Sci.* **14** 205–20

Konar M, Lin X, Ruddell B and Sivapalan M 2018 Scaling properties of food flow networks *PLoS One* **13** e0199498

Liang S, Wang H, Qu S, Feng T, Guan D, Fang H and Xu M 2016 Socioeconomic drivers of greenhouse gas emissions in the united states *Environ. Sci. Technol.* **50** 7535–45

Liang X-Z, Wu Y, Chambers R G, Schmoldt D L, Gao W, Liu C, Liu Y-A, Sun C and Kennedy J A 2017 Determining climate effects on US total agricultural productivity *Proc. Natl Acad. Sci.* **114** E2285–92

Lin X, Dang Q and Konar M 2014 A network analysis of food flows within the United States of America *Environ. Sci. Technol.* **48** 5439–47

Llordén G R 2017 Gamma Mixture Model Estimation with EM Algorithm MathWorks File Exchange (https://www.mathworks.com/matlabcentral/fileexchange/53028-gamma-mixture-model-estimation-with-em-algorithm)

Lobell D B, Schlenker W and Costa-Roberts J 2011 Climate trends and global crop production since 1980 *Science* **333** 616–20

Long S P, Marshall-Colon A and Zhu X-G 2015 Meeting the global food demand of the future by engineering crop photosynthesis and yield potential *Cell* **161** 56–66

Ludtke N, Panzeri S, Brown M, Broomhead D S, Knowles J, Montemurro M A and Kell D B 2007 Information-theoretic sensitivity analysis: a general method for credit assignment in complex networks *J. R. Soc. Interface* **5** 223–35

MacDonald G K, Brauman K A, Sun S, Carlson K M, Cassidy E S, Gerber J S and West P C 2015 Rethinking agricultural trade relationships in an era of globalization *BioScience* **65** 275–89

Marston L, Ao Y, Konar M, Mekonnen M and Hoekstra A Y 2018 High-resolution water footprints of production of the United States *Water Resour. Res.* **54** 2288–316

Mohri M, Rostamizadeh A and Talwalkar A 2012 *Foundations of Machine Learning* (Cambridge, MA: MIT Press)

Nesheim M, Oria M and Yih P 2015 *A Framework for Assessing Effects of the Food System* (Washington, DC: National Academies Press)

Oak Ridge National Laboratory 2011 County-to-county distance matrix (http://cta.ornl.gov/transnet/SkimTree.htm)

Oak Ridge National Laboratory 2015 Freight analysis framework version 4 (http://faf.ornl.gov/fafweb/)

Porkka M, Kummu M, Siebert S and Varis O 2013 From food insufficiency towards trade dependency: a historical analysis of global food availability *PLoS One* **8** e82,714

Rushforth R R and Ruddell B L 2018 A spatially detailed and economically complete blue water footprint of the United States *Hydrol. Earth Syst. Sci.* (https://doi.org/10.5194/hess-22-3007-2018)

Saltelli A, Tarantola S and Chan K-S 1999 A quantitative model-independent method for global sensitivity analysis of model output *Technometrics* **41** 39–56

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M and Tarantola S 2008 *Global Sensitivity Analysis: The Primer* (Chichester: Wiley)

Seekell D *et al* 2017 Resilience in the global food system *Environ. Res. Lett.* **12** 025010

Servadio J L and Convertino M 2018 Optimal information networks: application for data-driven integrated health in populations *Sci. Adv.* **4** e1701088

Smith T M, Goodkind A L, Kim T, Pelton R E O, Suh K and Schmitt J 2017 Subnational mobility and consumption-based environmental accounting of US corn in animal protein and ethanol supply chains *Proc. Natl Acad. Soc.* **11** E7891–9

US Bureau of Economic Analysis 2014 Input–Output Accounts Data (https://bea.gov/industry/io_annual.htm)

US Bureau of Economic Analysis 2017 Local Area Personal Income and Employment (https://bea.gov/iTable/itable.cfm?reqid=70&step=1&isuri=1&acrdn=7#reqid=70&step=1&isuri=1)

US Census Bureau 2015a 2012 Economic Census (http://census.gov/data.html)

US Census Bureau 2015b 2012 Commodity Flow Survey Public Use Microdata (https://census.gov/econ/cfs/pums.html)

US Census Bureau 2018 US Census Bureau USA Trade Database (https://usatrade.census.gov/index.php)

US Department of Agriculture 2014 National Agricultural Statistics Service Quick Stats (http://quickstats.nass.usda.gov)

USDA 2013 United States Department of Agriculture Economic Research Service (http://ers.usda.gov/data-products.aspx)

Venkatramanan S *et al* 2017 Towards robust models of food flows and their role in invasive species spread *2017 IEEE Int. Conf. on Big Data (BIGDATA)* (https://doi.org/10.1109/BigData.2017.8257955)

Vora N, Shah A, Bilec M M and Khanna V 2017 Food-energy-water nexus: quantifying embodied energy and ghg emissions from irrigation through virtual water transfers in food trade *ACS Sustain. Chem. Eng.* **5** 2119–28

Walker S and Duncan D 1967 Estimation of the probability of an event as a function of several independent variables *Biometrika* **54** 167–78

Wang G, Hao J, Ma J and Jiang H 2011 A comparative assessment of ensemble learning for credit scoring *Expert Syst. Appl.* **38** 223–30

Wang R, Zimmerman J B, Wang C, Vivanco D F and Hertwich E G 2017 Freshwater vulnerability beyond local water stress: heterogeneous effects of water-electricity nexus across the continental United States *Environ. Sci. Technol.* **51** 9899–910

Weber C L and Matthews H S 2008 Food-miles and the relative climate impacts of food choices in the United States *Environ. Sci. Technol.* **42** 3508–13

Xu M, Allenby B R and Crittenden J C 2011 Interconnectedness and resilience of the US economy *Adv. Complex Syst.* **14** 649–72