Modeling U.S. cattle movements until the cows come home: Who ships to whom and how many?

Stefan Sellman a,b, Lindsay M. Beck-Johnson b, Clayton Hallman c, Ryan S. Miller c, Katharine A. Owers Bonner b, Katie Portacci c, Colleen T. Webb b, Tom Lindström a

aDepartment of Physics, Chemistry and Biology, Division of Ecological and Environmental Modeling, Linköping University, Linköping, Sweden
bDepartment of Biology, Colorado State University, Fort Collins, CO, USA
cUSDA APHIS Veterinary Services, Center for Epidemiology and Animal Health, Fort Collins, CO, USA

A B S T R A C T

Livestock movements between agricultural premises is an important pathway for the spread of infectious disease. Data providing details about the origin and destination of shipments, as well as information about the shipment size is an important component of computer models used to formulate mitigation strategies and design surveillance programs. The United States (U.S.) currently lacks a comprehensive database of farm animal shipments, which hinders such efforts. With the U.S. Animal Movement Model (USAMM), earlier work has successfully scaled up from limited data based on interstate certificates of veterinary inspection (CVI) to comprehensive county-level shipment networks at the national scale. In this work, we present three major improvements to earlier versions of USAMM: (1) increased resolution of the model and simulated networks to the level of individual premises; (2) predictions of shipment sizes; (3) taking into account the types and herd sizes of the premises. We fitted parameters in a Bayesian framework to two sets of CVI data consisting of sub-samples of one year’s between-state beef and dairy shipments. Through posterior predictive simulation, we then created 1,000 synthetic beef and dairy networks, which we make publicly available to support livestock disease modeling. The simulated networks were validated against summary statistics of the training data as well as out-of-sample CVI data from subsequent years. This new development opens up the possibility of using USAMM in a broader spectrum of applications where information about shipment size and premises identity is necessary and gives novel insights into the U.S. cattle shipment network.

1. Introduction

Transportation of livestock between herds is a common and necessary part of modern agricultural practices. But the act of moving animals also pose the risk of facilitating the spread of infectious disease by introducing infected animals and fomites to susceptible herds (Fèvre et al., 2006). Therefore, the transportation of live animal has often been implicated as an important epidemiological factor for a wide variety of important livestock diseases such as foot-and-mouth disease (FMD Gibbens et al., 2001; Muroga et al., 2012), bovine tuberculosis (bTB, Palisson et al., 2016; Bessell et al., 2012; Byrne et al., 2017; Okafor et al., 2011; Gilbert et al., 2005), scrapie (Gubbins, 2005), avian influenza (Guinat et al., 2020; Suarez, 2010), porcine reproductive and respiratory syndrome (VanderWaal et al., 2020) and porcine epidemic diarrhea (Machado et al., 2019). Not only do these examples demonstrate that animal movements are important for the transmission of livestock disease in a variety of host species, but notably also for diseases with very different transmission dynamics. For instance, FMD is highly contagious, spreads rapidly and has the potential to develop into widespread epidemics in a matter of weeks or months (Keeling et al., 2001). In contrast, bTB has a slow development with an infectious period of many months, and can remain undetectable in infected animals for long periods, making it difficult to diagnose and detect (Álvarez et al., 2014; O’Hare et al., 2014).

The costs associated with infectious livestock diseases are high, both economically and in terms of animal and human welfare (Tomley and Shirley, 2009). This motivates efforts to formulate surveillance and control strategies aimed to minimize the risk of outbreaks, as well as to control and reduce the extent of outbreaks of they do occur. Such efforts often calls for detailed spatial information about the system, and given the importance of live animal movements for the spread of disease a thorough understanding of the network of livestock shipments between herds is critical (Rautureau et al., 2011; Brooks-Pollock et al.,...
2015). Such information can assist in the choice of high-risk premises to monitor for early warnings, inform where to concentrate surveillance efforts, promote detection of high-risk contacts, and allow evaluation of control programs using computer simulations of outbreaks (Brooks-Pollock et al., 2015; Gorsich et al., 2018; Kinsley et al., 2019; Gates and Woolhouse, 2015; van Andel et al., 2021).

For these reasons, many governments require livestock movements to be registered in national databases where the information can quickly be accessed if infection is detected in a population (Caporale et al., 2001). Examples of countries where registration of animals and their movements are mandatory for at least one livestock species for the purpose of enabling such movement tracing include (but are not limited to) the United Kingdom (Green and Kao, 2007), Japan (Sugiura and Onodera, 2008), Argentina (Alarcon et al., 2020), Uruguay (VanderWaal et al., 2016), and all member states of the European union (Amendrup and Füssel, 2001). Despite previous attempts to establish such a national database, the U.S. currently lacks a comprehensive system to trace and identify individual animals, animal holdings or livestock shipments. In 2004 the National Animal Identification System (NAIS) was introduced in the U.S. aiming to set up a mandatory register of all animal holding facilities and assigning identification numbers to livestock and batches of livestock which would enable tracing of movements between premises (Skaggs, 2011). However, the system was faced with heavy criticism from livestock producers due to concerns regarding privacy, confidentiality, and economic costs, and the program was canceled in 2010 (Anderson, 2010). In 2018, another initiative, CattleTrace, was launched, later turning into the privately owned corporation U.S. CattleTrace, offering a voluntary program for electronic identification of individual cattle in order to facilitate traceability of livestock and animal products (Shear and Pendell, 2020, https://www.uscattletrace.org/). With enough participants, U.S. CattleTrace has the potential to be a useful resource for livestock disease control and prevention by enabling rapid tracing of infected animals in the case of an outbreak. However, private data collected and maintained by the industry itself may be inaccessible for use in broader applications within epidemiological research (e.g. for modeling and designing effective control strategies) if the participants’ right to confidentiality has been assured by the owner of the data. Further, the non-mandatory nature of any such program may present an obstacle to statistical inference through selection bias if participants are not representative of the entire population. The only set of U.S. cattle shipment data that is mandated by federal law to be collected by official authorities consists of records of interstate certificates of veterinary inspections (CVIs). These are official documents issued by accredited state veterinarians when cattle and other livestock species are being shipped across state borders for non-slaughter purposes (C.F.R, 2021). The main purpose of the CVIs is to certify that shipped animals are healthy, but they include information which make them useful for shipment modeling, such as the locations, shipment sizes (i.e. number of animals), and, occasionally, types of the origin and destination premises (Buhnerkempe et al., 2013). As the CVIs are required by federal law, they represent the only data set with complete geographical coverage of the entire conterminous U.S. At the same time, since CVIs are required only when shipping animals across state borders, they also represent a biased sample of the U.S. cattle trade network, making these records unsuitable to directly inform modeling and policy.

A component that is lacking from all previously mentioned models, however, is the prediction of shipment size. The number of animals that are moved between two premises will have an impact on the probability that transmission between the two herds occurs as the number of infected animals that are shipped will be directly affected by the size of the shipment in addition to the prevalence within the origin herd. If information regarding shipment size is unavailable when modeling between-herd dynamics of infectious diseases, simplifying assumptions about the contribution of shipments to the disease transmission has to be made. This is less of an issue for fast-moving diseases such as FMD, where entire herds are commonly modeled as either infected or susceptible (e.g. Keeling et al., 2001), but may be of more importance when modeling diseases with slow transmission such as bTB, where often only a fraction of a herd is infected at a given time and detailed within-herd modeling is employed (e.g. Brooks-Pollock et al., 2014 and Birch et al., 2018). For the modeling of such situations in the U.S., we believe a model of the shipment network that is able to predict shipment size will be useful, but such a model would have to include premises-level attributes in order to be realistic. For instance, it is reasonable to expect that shipment size depends to some extent on the herd sizes of the premises’ involved in the shipment and that different types of premises will display different shipment patterns (Bates et al., 2001). For instance, markets act as hubs in networks of livestock movements with higher in- and out-degree than other kinds of operations (Robinson and Chrisley, 2007; Vidondo and Voelkl, 2018), and feedlots play an important role in the U.S. beef production system, concentrating animals from other kinds of operations (Shields and Mathews, 2003). Given these differences, it is conceivable that shipment sizes also varies with the types of premises involved.

As mentioned, CVIs include the shipment size and sometimes types of the involved premises, in addition to their location, but this information has not been utilized in any of the previous models. Here we present a substantially improved version of USAMM for cattle that takes this information into account in order to model very detailed shipment networks at the resolution of individual premises that include predictions about shipment size. The increase in level of detail compared to previous versions is introduced in order to enable the use of premises-level attributes, such as herd size and premises type as predictors in the model, which we believe is necessary in order to realistically model shipment size. We thoroughly describe the new model,
fit it to two sets of cattle CVI data describing different commodities (beef and dairy) and construct 1000 synthetic shipment networks for each commodity through posterior predictive simulations. Using these synthetic networks we perform within- and out-of-sample validation to evaluate if the model is capable of simulating networks that are comparable to the CVI data sets in terms of the spatial distribution of shipments and shipment size distributions. Finally, in order to facilitate further studies involving the U.S. cattle industry, we make the synthetic beef and dairy shipment networks available in their raw form at http://dx.doi.org/10.25675/10217/234115 and in an interactive format at https://usamm-gen-net.shinyapps.io/usamm-gen-net/.

2. Materials and methods

2.1. Overview

We constructed a hierarchical Bayesian model that was fit to two separate incomplete, but systematically collected, CVI data sets describing the U.S. interstate beef and dairy shipments during the year 2009, respectively. The model was built around the assumption that the number of shipments between each unique pair of premises, given by a separate demography data set, during a single day can be described as a Poisson distributed random variable (RV), and hence each unique pair was associated with a daily shipment rate. This shipment rate was modeled using various covariates: the herd size of the premises, the distance between the premises, measures related to the cattle industry of the origin and destination counties, and in- and outflow weights of the origin and destination states. In addition to the number of shipments between premises, the size of each shipment was modeled as a RV following a premises type-specific probability distribution which included the premises' herd sizes as covariates. The importance of these covariates for determining the shipment rate and shipment size was controlled by a large number of model parameters, many of which were defined separately for different premises types (farm, feedlot, market). Furthermore, in order to capture seasonally dependent shipment patterns, most parameters were also defined separately for the four quarters of the year. The model parameters were estimated iteratively using a Markov chain Monte Carlo (MCMC) approach with a combination of the Metropolis–Hastings algorithm and Gibbs-sampling, coded in C++ and tailored specifically for this problem and data. In order to include herd size and premises type as covariates in the model, the identity of the origin and destination premises was needed, but only the origin and destination counties were available from the shipment data. Therefore, the exact premises identities in relation to the demography data were also estimated as additional parameters by sampling them every model iteration, one by one, conditionally on all the other current model parameters, from the respective counties’ premises demography. The beef and dairy networks were analyzed separately and independently, and from each of the two resulting posterior parameter distributions 1,000 samples were drawn and used for posterior predictive simulation of as many complete networks of the number of shipments between each county containing premises, a centroid coordinate and total area was required. These were obtained from publicly available shapefiles published by the U.S. Census Bureau (https://www.census.gov/programs-surveys/geography/geographies.html).

2.2. Data

Three sets of data were used as input to the model: county-level premises demography, CVI shipment data, and county-level cattle industry covariates. In the sections below we outline each of these three data sets. Additionally, for each county containing premises, a centroid coordinate and total area was required. These were obtained from publicly available shapefiles published by the U.S. Census Bureau. (https://www.census.gov/programs-surveys/geography/geographies.html).

2.2.1. Premises demography

In order to capture behaviors and dynamics specific to separate sectors of the cattle rearing industry the model differentiates between different types of premises by associating each with its own set of model parameters. The types included were beef farms, dairy farms, feedlots and markets. The model works at the level of individual premises, assigning weights to each premises within a given county based on their herd size, and hence information about the numbers of premises of different size in each county is required as input. Although no complete demography data exists for U.S. cattle premises, The National Agricultural Statistical Survey (USDA, 2014a) census provides information about the number of beef farms, dairy farms, and feedlots in a fixed set of discrete size bins by premises type at the level of counties. However, if there are fewer than three premises in a single county or if the premises size distribution within the county is such that the identity of one or more premises could be easily determined (USDA, 2014b), NASS withholds information for the county to protect the privacy of these premises. To work around this we used synthetic farm and feedlot populations generated with the Farm Location and Animal Population Simulator (FLAPS, Burdett et al., 2015) as input to the model, rather than the raw NASS survey data. FLAPS simulates county-level premises size distributions based on the information in the 2012 NASS census data, imputing the premises distribution in counties where data are withheld. Basing the demography data used in the model on the 2007 NASS census would have been slightly closer in time to the available shipment training data (2009, see below), but no such FLAPS version is available. Informed by the NASS county-level binned size distributions, FLAPS also simulates an exact herd size of individual farms and feedlots. However, in order to reduce the reliance on imputed data, we used the original NASS survey data where available, and only relied on sizes simulated by FLAPS for the censored counties for which size information would otherwise be unavailable. In practise, this was achieved by reassigning all premises sizes simulated by FLAPS back to the size bins defined by NASS as detailed in Table 1. The binning of the premises sizes was also a necessity for computational feasibility. The largest NASS size bin is defined as ≥500 cattle, but many feedlots are significantly larger than that. Therefore, for feedlots, we added an upper limit to the largest NASS size bin at 999 cattle and introduced an additional larger bin, defined as ≥1000 cattle. Separate realizations of the demography were simulated with FLAPS for beef farms, dairy farms and feedlots; each informed by a subset of the premises recorded in the NASS census. In the census, “any place from which $1000 or more of agricultural products were produced and sold, or normally would have been sold, during the census year” (USDA, 2014b) is included. All such places with beef or dairy cattle present or with sales thereof during the year of the census, but excluding those with “Cattle on feed” in the census, were used to inform the beef and dairy farm FLAPS realizations, while only those with “Cattle on feed” were used for the feedlot realization. By consequence, our definition of farm used here includes all premises with cattle present that are not feedlots (i.e. cow–calf operations, backgrounders, ranches, stockers, etc.).

Live cattle markets and auctions is an important element of the U.S. cattle industry and were included in the model as a separate premises type which we refer to simply as market. The NASS census (and consequently FLAPS) does not provide information regarding cattle market demography, which therefore had to be informed by other sources. However, a list of permanent cattle markets in counties across the U.S. has been compiled from publicly available data sets by Carroll and Bansal (2015). This list was combined with the average annual market sales volumes of 266 counties from historic market sales data obtained from the USDA Agricultural Marketing Service (AMS). For counties where such historic market sales data was unavailable, a spatial simultaneous autoregressive lag model (Bivand et al., 2013) was used to estimate the sales volumes for the other markets, taking into account the market volumes in the surrounding counties and using the county total cattle sales (USDA, 2014a) as a covariate. The volume
of each individual market was calculated as the total county-level annual market sales estimate divided by the number of markets in the county as given by the market list. Thus, in contrast to the size of farms and feedlots, the size of each market was defined as a yearly turn-over of animals. As the NASS census does not contain market information, there are no predefined size bins for markets and they were therefore defined subjectively. Further, no distinction was made between turn-over of beef or dairy cattle, as that information is not available.

The final farm and feedlot demographies consisted of 723,679 beef farms, 63,984 dairy farms, and 26,585 feedlots. The final market demography consisted of 1624 markets in 1131 different counties, out of which 361 counties had more than one market. We represent these four populations as four sets that each consist of all individual premises of each corresponding type: \( P_{\text{Farm}} \), \( P_{\text{Dairy}} \), \( P_{\text{Feed}} \) and \( P_{\text{Market}} \). With these type-specific sets as building blocks we define two additional sets of premises that correspond to the demographies of the beef and dairy industries: \( P_{\text{Beef}} = (P_{\text{Farm}} \cup P_{\text{Feed}} \cup P_{\text{Market}}) \) and \( P_{\text{Dairy}} = (P_{\text{Farm}} \cup P_{\text{Feed}} \cup P_{\text{Market}}) \). For every premises \( i \) of premises type \( t_i \), in these sets we define the herd size \( h_i \) as the midpoint of the size interval (rounded up) to which the premises belonged. Since the largest categories lacked upper limits, there was no midpoint to assign for premises in that interval. For these cases, we defined \( h_i \) to be the average size of all premises of type \( t_i \) in the largest size bin as simulated by FLAPS, rounded to the nearest multiple of 100 for feedlots and the nearest multiple of 10,000 for markets. Consequently, \( h_i \) for premises belonging to the largest size bins were defined as 1200 for farms, 6500 for feedlots, and 140,000 for markets (Table 1).

Table 1

<table>
<thead>
<tr>
<th>Premsises type</th>
<th>Bin lower limit</th>
<th>Bin upper limit</th>
<th>( h_i )</th>
<th>Number of premises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef farms</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>93,900</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>19</td>
<td>15</td>
<td>135,341</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>49</td>
<td>35</td>
<td>184,397</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>99</td>
<td>75</td>
<td>165,062</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>199</td>
<td>150</td>
<td>72,166</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>499</td>
<td>350</td>
<td>46,824</td>
</tr>
<tr>
<td></td>
<td>( \geq 500 )</td>
<td>( \geq 1200 )</td>
<td></td>
<td>25,989</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>723,679</td>
</tr>
<tr>
<td>Dairy farms</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>15,532</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>19</td>
<td>15</td>
<td>1139</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>49</td>
<td>35</td>
<td>6531</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>99</td>
<td>75</td>
<td>13,048</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>199</td>
<td>150</td>
<td>14,373</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>499</td>
<td>350</td>
<td>7910</td>
</tr>
<tr>
<td></td>
<td>( \geq 500 )</td>
<td>( \geq 1200 )</td>
<td></td>
<td>5451</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>63,984</td>
</tr>
<tr>
<td>Feedlots</td>
<td>0</td>
<td>1,999</td>
<td>5,000</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>19,999</td>
<td>15,000</td>
<td>718</td>
</tr>
<tr>
<td></td>
<td>20,000</td>
<td>39,999</td>
<td>30,000</td>
<td>557</td>
</tr>
<tr>
<td></td>
<td>40,000</td>
<td>59,999</td>
<td>50,000</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>60,000</td>
<td>79,999</td>
<td>70,000</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>80,000</td>
<td>99,999</td>
<td>90,000</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>( \geq 100,000 )</td>
<td>( \geq 160,000 )</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>1624</td>
</tr>
</tbody>
</table>

2.2.2. Shipment data

The shipment data used to estimate the model parameters was based on CVIs for outgoing shipments from 47 out of the 48 conterminous U.S. states (New Jersey did not provide data). The data set was constructed from a systematic 10% sample of CVI paper records collected in 2009 and is described in depth in Buhnerkempe et al. (2013). CVIs are not required for shipments of animals bound for slaughter, so the analysis includes no slaughter shipments. Information included in the data relevant for this study was counties of origin and destination, number of animals shipped, date on which the shipment occurred, whether the shipment contained beef or dairy animals, and the names and addresses of the origin and destination operations. From the original sample of 19,817 CVIs collected by Buhnerkempe et al. (2013) we removed 14 entries due to lack of information on the number of animals. A further 86 entries were removed due to missing date of shipment and 642 entries where the destination state was the same as the origin were removed. This large number of intrastate shipments is due to the fact that CVIs are sometimes issued for other reasons than inter-state travel (Buhnerkempe et al., 2013). Finally, 187 shipments were removed due to the origin or destination county having no premises according to the premises demography data, and 45 shipments were removed because there were no premises large enough to of suitable size in either county (see Section 2.3.2). The original proportion of observed outgoing shipments of 10% for each state was adjusted for the removed shipments to get a specific proportion \( z_s \) reflecting the corrected sample size for each state \( s \). The final data set consisted of 18,843 CVI entries divided into 16,193 beef and 2650 dairy shipments with information about the origin and destination counties and states, the number of animals shipped, and the shipping date.

The original CVI records do not include explicit information about the origin and destination premises types. However, the names and addresses of the premises involved in the shipment are generally available. By matching this information to keywords indicating premises type and known market names, we assigned a premises type to both the sending and receiving premises of the shipment (see Appendix S1 for further details) as either feedlot or market where possible. Keywords that exclusively indicated that a premises was a farm were not identified. The relevant information was not available for all shipments, and many names did not provide a reliable match. Designating a specific origin or destination type to a shipment in this way meant that the model was restricted to only assign premises of feedlot or market type as sender or receiver of that particular shipment. If no premises type could be determined for one of the premises involved in the shipment, the model was allowed to assign the shipment to a premises of any type. Origin and destination types were assigned independently of each other and a shipment could be assigned either, both, or none depending on the information available for the given shipment. For the beef shipments, 5986 origin and 464 destination premises, were identified as markets respectively, while 123 origin and 2937 destination premises were identified as feedlots. For the dairy shipments, 282 origin and 102 destination premises were assigned as markets, while 59 and 103 origin and destination premises were identified as feedlots.

Two additional sets of shipment data based on CVIs for the years 2010 and 2011 were collected and thoroughly analyzed in Gorsich et al. (2016). These data set cover a 10% sample from a subset of eight origin states and were used in this study for the purpose of out-of-sample model validation.

The shipment data sets used in this study are relatively old, and could be considered out-of-date. Comprehensive statistics that would allow rigorous assessment of whether it is so or not are scarce, but NASS performs a yearly survey of the U.S. cattle inventory where the number of operations, animals, and inbound shipments are estimated based on a subset of the U.S. cattle farms (https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Cattle_Inventory/index.php). Using this source, we analyzed the cattle industry for any recent general trends that may impact the credibility of the results.
2.2.3. Cattle industry covariates

Industry covariates used in the model to explain county-level variation of premises-to-premises shipment rates were selected from 13 different potential cattle industry related measures. Eleven of these were data items from the NASS agricultural census (USDA, 2014a) identified by the authors as potentially relevant for beef and dairy production; one was a measure of closeness to slaughter capacity (slaughter connectivity, see below); and one was the number of markets in the county from the market data. A first selection was made based on the predictive power \((R^2)\) of each potential covariate considered individually as a predictor in a univariable linear model of the county-level in- and out-degree of the CVI data. Based on the results of that analysis, groups of covariates were selected and used as predictors jointly in a multivariable linear model and a final selection was made based on the Akaike information criterion (AIC). The combination of covariates with the best explanatory power consisted of the number of operations with sales in the last year (OS), slaughter connectivity (SC), and total number of cattle sold (TS). Additionally, to disentangle the effect of individual premises’ herd size on shipment rate from the effect of the county-level herd size and to control for the general size of the cattle industry of the county we also added the total number of cattle (NC) of each county as a fourth covariate. For the eleven potential covariates from NASS, data was preferentially obtained from the 2012 census, which is closest in time to the shipment data out of all available censuses. However, as mentioned in Section 2.2.1, NASS censors some county-level data in order to avoid potential privacy concerns (USDA, 2014b). When this was the case for a county in the 2012 census, the 2007, 2002 and 1997 censuses were used to fill in any censored values by using the most recent year from which they were available. Across all combinations of counties and NASS covariates there were 2217 (6.5%) data points that were censored in the 2012 NASS census, and 414 (1.2%) data points that were censored in all of the four censuses. These remaining data points were imputed through a random forest linear model using rfimp() from the randomForest package in R (Liaw and Wiener, 2002). In short, random forests are classification algorithms that use a set of predictors to group observations based on similarity. When the response variable (the classification) is known, they can be used to find good estimates to missing values among the predictors. Here we imputed the missing values in the county-level NASS categories by using them as predictors and the state to which the counties belong as a separate response variable. For a list of the original 11 NASS predictors and the number of censored values, see Table S1.

We expected to find some relationship between the number of animals slaughtered in a region and the non-slaughter shipment patterns. However, since slaughter does not necessarily take place in the same county that the animals were reared in, a county without a slaughter facility can still be an important link in the chain of movements that lead up to slaughter if there are large slaughter facilities close by. Therefore we constructed a metric to measure the amount of slaughter capacity available to each county rather than relying directly on the number of animals slaughtered within the counties. The metric, referred to as slaughter connectivity, \(I_c\), was calculated for each county \(c\) as the distance-weighted sum of the annual slaughter volume in head of cattle of over all other counties,

\[
I_c = \sum_{y \neq c} \frac{d_{y,c}^2 e^{-d_{y,c}/125}}{e^{-d_{y,c}/125}}
\]

where \(I_c\) is the slaughter volume in head of county \(x\), \(c\) is the set of all counties, and \(d_{y,c}\) is the distance between the centroids of \(c\) and \(x\) in km.

The distances were normalized by an estimate of the average distance of over all other counties, distance-weighted sum of the annual slaughter volume in head of cattle \(\hat{I}_c\) available to each county rather than relying directly on the number animals slaughtered in a region and the non-slaughter shipment patterns. Therefore they can be used to find good estimates to missing values among the similar. When the response variable (the classification) is known, \(\eta_i\) is the baseline rate which depends solely on the premises types \(i\) and \(j\). Second, this baseline rate is given a seasonal dependence by multiplication with the factor \(\zeta_{t/q}\) that scales the baseline rate depending on the quarter of

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s)</td>
<td>State.</td>
</tr>
<tr>
<td>(c)</td>
<td>Counties.</td>
</tr>
<tr>
<td>(x)</td>
<td>Precinct.</td>
</tr>
<tr>
<td>(w)</td>
<td>Shipments.</td>
</tr>
<tr>
<td>(q)</td>
<td>Quarter of the year.</td>
</tr>
<tr>
<td>(P)</td>
<td>The set of all premises.</td>
</tr>
<tr>
<td>(M)</td>
<td>The set of all shipments.</td>
</tr>
<tr>
<td>(d_{ij})</td>
<td>Distance (m) between premises (i) and (j).</td>
</tr>
<tr>
<td>(h_i)</td>
<td>Herd size of the premises (i) (yearly volume for markets).</td>
</tr>
<tr>
<td>(v_w)</td>
<td>Shipments volume (number of animals) of shipment (w).</td>
</tr>
<tr>
<td>(u_{ij})</td>
<td>Number of shipments from (i) to (j) in quarter (q).</td>
</tr>
<tr>
<td>(r_c)</td>
<td>County level industry covariates of county (c).</td>
</tr>
<tr>
<td>(z_i)</td>
<td>Proportion of interstate shipments from state (i) observed in the data.</td>
</tr>
</tbody>
</table>

2.3. Statistical model

In summary, the number of shipments between each unique pair of premises \(i\) and \(j\) in each quarter \(q\), is modeled as a Poisson distributed random variable, \(u_{ij,q}\). For each shipment, \(m\), the shipment size, \(v_m\), is modeled as a random variable which follows a gamma-Poisson, or a beta-binomial distribution depending of the types of premises that are involved in the shipment. These distributions are governed by a set of model parameters referred to as \(\theta\). The model assumes that the identities, in the form of herd size and type, of the origin and destination premises of each shipment are known. However, as mentioned in the previous sections, the premises demography is only known to the level of the type-specific size-class distribution within each county. Therefore, even though the CVI data sometimes contain address information, it is not possible to connect shipments to particular premises from the data directly. Instead, the model includes a component that probabilistically assigns an origin and a destination premises to each shipment from the available premises population of the respective origin and destination counties conditional on \(\theta\). In this way, the identities of the origin and destination premises are estimated alongside the model parameters.

We begin this section with the statistical approach regarding how \(u_{ij,q}\), and then \(v_m\) were modeled, specify the likelihood function and describe the choice of model priors. We reserve the use of Greek letters to parameters estimated in the model, while other variables and functions are denoted by Latin letters. Tables 2 and 3 provides a summary of what the various symbols and parameters represent.

2.3.1. Number of shipments

The model is constructed around the assumption that the number of shipments, \(u_{ij,q}\), between every unique combination of origin, \(i\), and destination, \(j\), cattle premises in quarter \(q\) \(\in \{1,2,3,4\}\) can be described by a Poisson process with rate \(\eta_{ij,q}\),

\[
u_{ij,q} \sim \text{Pois}(\eta_{ij,q}).
\]

where \(j\) and \(q\) are part of either \(\mathbf{P}^{\text{Beef}}\) or \(\mathbf{P}^{\text{Dairy}}\) depending on what industry is being modeled. We model this premises-to-premises shipment rate as a function of factors at the level of the focal premises pair and their respective origin and destination premises types \((i, j, t)\), counties \((c, c)\), and states \((s, s)\). The rate is defined as

\[
u_{ij,q} = \eta_{ij,q} \zeta_{t/q} \delta_{i,j,q} \delta_{t/q} \sigma_{i,q} \gamma_{j,q} K(a_{i,j,q}, b_{i,j,q}).
\]

For parameters at the level premises we first define \(\eta_{ij,q}\) as a baseline rate that depends solely on the premises types \(i\) and \(j\). Second, this baseline rate is given a seasonal dependence by multiplication with the factor \(\zeta_{t/q}\) that scales the baseline rate depending on the quarter of
the year, \(q\). Third, we let the herd sizes of the two premises play a role via the factor \(f_{ijq}\), adjusting each premises’ contribution according to a function of its herd size:

\[
f_{ijq} = \exp \left( \hat{h}_i \phi_{ijq}^\text{out} + \hat{h}_j \phi_{ijq}^\text{in} \right),
\]

where \(\hat{h}_i\) is the logarithm of the normalized herd size of a premise and \(\phi\) are scaling exponents that depends on both premises type and direction of the shipment (i.e. if the premises in question is receiving or sending the shipment). We normalize the herd sizes by dividing by the average herd size of all premises of the same type:

\[
\hat{h}_i = \ln \left( \frac{h_i}{\bar{h}_i} \right) = \ln \left( \frac{\left| P_i \right|}{\sum_{j \in P^x} h_j} \right),
\]

where \(h_i\) is the unnormalized binned herd size of premises \(i\), \(\bar{h}_i\) is the premises type of \(x\) and \(P^x\) is the set of all premises of type \(x\) (i.e. \(P^\text{Fall}\), \(P^\text{Winter}\), \(P^\text{Cold}\) or \(P^\text{Market}\)). This normalization was used to reduce the absolute differences between the largest and the smallest herds while retaining the relative differences, improving the numerical stability of the model and making the parameter estimation step more efficient. It also aids in interpreting of the model parameters as the weight of a premises average size will always be equal to unity in Eq. (4), meaning that the parameters \(\phi^\text{out}\) and \(\phi^\text{in}\) control to what extent a deviation from the average herd size will affect the shipment rate.

Next, we include an effect of features related to the cattle industry via the county-level covariates. Denoting the counties of \(i\) and \(j\) as \(c_i\) and \(c_j\), and, (departing from the usual index notation for clarity) the vector of covariates associated with these counties as \(r(c_i)\) and \(r(c_j)\), they affect the premises-to-premises shipment rate via the factors \(g^\text{out}\) and \(g^\text{in}\) given by equations

\[
g^\text{out} = \exp \sum_{k=1}^{n} k \gamma_k \cdot r(c_k)_i,
\]

and

\[
g^\text{in} = \exp \sum_{k=1}^{n} j \gamma_k \cdot r(c_k)_i.
\]

Here, \(n\) is the number of covariates included in the model, and \(\gamma^\text{out}\) and \(\gamma^\text{in}\) are weights associated with covariate \(r\). The country level covariates are standardized by shifting the mean across counties to be zero and the standard deviation to be 0.5.

Finally, at the state level there are three factors that affect the shipment rate between \(i\) and \(j\). Two of these are simple random effects: \(\omega^\text{out}\) of the origin state \(s_i\) that \(i\) belongs to, and \(\omega^\text{in}\) of the destination state \(s_j\) that \(j\) belongs to. The third is a factor that introduces an effect of distance on the shipment rate according to the monotonically decreasing function:

\[
K(d_{ij}, a_{ij}, b_{ij}) = \frac{1}{1 + \left( \frac{d_{ij}}{a_{ij}} \right)^{b_{ij}}},
\]

where \(d_{ij}\) is a measure of distance between \(i\) and \(j\), and the kernel shape and scale parameters, \(a_{ij}\) and \(b_{ij}\), are specific to the state of origin, \(s_i\). This specific form of the kernel function was determined to perform best out of three potential candidate functions for cattle shipments in a previous work (Brommesson et al., 2021). Since no exact farm locations are available, we define \(d_{ij}\) as a binned distance between the centroids of the respective counties of \(i\) and \(j\) and use that for our measure of \(d_{ij}\). This limits the number of unique distances the function needs to be evaluated for, reducing computational complexity. With \(d_{ij}^\text{cnt}\) as the centroid-to-centroid distance, binning was achieved by rounding \(d_{ij}^\text{cnt}\) to the nearest multiple of \(10^5\) if \(d_{ij}^\text{cnt} < 10^5\) m; to the nearest multiple of \(10^4\) if \(10^5 \leq d_{ij}^\text{cnt} < 10^6\) m; and to the nearest multiple of \(10^3\) if \(d_{ij}^\text{cnt} \geq 10^6\) m.

### 2.3.2. Shipment size

We model the shipment size \(v_m\) of a given shipment \(m\) between origin premises \(i_m\) and destination premises \(j_m\) in quarter \(q_m\) in two different ways depending on the types of the premises that are involved in the shipment. As a baseline assumption we let \(v_m\) be relative to the herd size of the origin premises, \(h_i\), by fitting a parameter that controls what proportion of the herd is sent. However, since market sizes are estimated from yearly sales and is not directly comparable to the herd size of a farm or feedlot, we model any shipment that involves

| Table 3 |
| List of model parameters and related prior distributions. The parameters are given with subscripts indicating origin premises type (\(i\)), destination premises type (\(j\)), quarter (\(q\)), origin state (\(s\)) and destination state (\(t\)). The rightmost column summarizes the prior distribution used for each parameter; if the distribution is defined as having 95% of the density between two values (percentiles), these two values are indicated as \(\nu_{5,95}\) and \(\nu_{5,95}\). |

<table>
<thead>
<tr>
<th>Parameters related to shipment rate ($v_m$)</th>
<th>Prior distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\phi_{ij}^\text{out}, \phi_{ij}^\text{in})</td>
<td>(\mathcal{N}(\nu_{5,95} = -0.5, \nu_{5,95} = 1.5))</td>
</tr>
<tr>
<td>(\beta_{ij})</td>
<td>(\Gamma(\text{shape} = 1 \times 10^{-3}, \text{rate} = 1 \times 10^{-3}))</td>
</tr>
<tr>
<td>(\alpha_{ij})</td>
<td>(\Gamma(\text{shape} = 0.1, \nu_{5,95} = 10.0))</td>
</tr>
<tr>
<td>(\gamma_{ij})</td>
<td>(\mathcal{LN}(\text{mean} = \mu^\gamma, \text{CV} = 4^\gamma))</td>
</tr>
<tr>
<td>(\omega_{ij}^\text{out})</td>
<td>(\Gamma(\text{shape} = \omega^\text{out}, \text{mean} = 1.0))</td>
</tr>
<tr>
<td>(\omega_{ij}^\text{in})</td>
<td>(\Gamma(\text{shape} = \omega^\text{in}, \text{mean} = 1.0))</td>
</tr>
<tr>
<td>(\gamma_{ij}^\text{cnt})</td>
<td>(\mathcal{N}(\nu_{5,95} = -2.0, \nu_{5,95} = 2.0))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters related to shipment size ($v_s$)</th>
<th>Prior distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\rho_{ij})</td>
<td>(\mathcal{U}(0,1))</td>
</tr>
<tr>
<td>(\nu_{ij})</td>
<td>(\mathcal{LN}(\nu_{5,95} = 2.0, \nu_{5,95} = 200))</td>
</tr>
<tr>
<td>(\xi)</td>
<td>(\mathcal{LN}(\nu_{5,95} = 12.1, \nu_{5,95} = 1210))</td>
</tr>
<tr>
<td>(\zeta)</td>
<td>(\mathcal{LN}(\nu_{5,95} = 0.1, \nu_{5,95} = 100))</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters related to priors and hyperpriors.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu^\text{prior})</td>
<td>(\mathcal{LN}(\nu_{5,95} = 1.0 \times 10^6, \nu_{5,95} = 4.0 \times 10^6))</td>
</tr>
<tr>
<td>(\xi^\text{prior})</td>
<td>(\mathcal{LN}(\nu_{5,95} = 0.365, \nu_{5,95} = 1.724))</td>
</tr>
<tr>
<td>(\omega^\text{prior})</td>
<td>(\mathcal{LN}(\nu_{5,95} = 2.0, \nu_{5,95} = 1000))</td>
</tr>
<tr>
<td>(\nu^\text{prior})</td>
<td>(\mathcal{LN}(\nu_{5,95} = 0.365, \nu_{5,95} = 1.724))</td>
</tr>
<tr>
<td>(\gamma^\text{prior})</td>
<td>(\mathcal{HC}(\text{scale} = 1.0))</td>
</tr>
<tr>
<td>(\delta^\text{prior})</td>
<td>(\mathcal{HC}(\text{scale} = 1.0))</td>
</tr>
</tbody>
</table>

\(\mathcal{N}\)—normally distributed, \(\Gamma\)—gamma distributed, \(\mathcal{LN}\)—lognormally distributed, \(\mathcal{U}\)—uniformly distributed, \(\mathcal{HC}\)—half-Cauchy distributed, CV—coefficient of variation.
a market independently of the market’s volume. Thus, when a market is not the origin of a shipment, we model \( v_m \) as a beta-binomial random variable with \( h_m \) as the number of trials (Appendix S2). When the origin premise is a market, but the destination premise is either a farm or a feedlot, we assume that the size of the purchase will be relative to the herd size of the destination herd size and let \( v_m \) be a beta binomial random variable with \( h_m \) as the number of trials. When both \( i \) and \( j \) are markets we assume that \( v_m \) follows a gamma-Poisson distribution (Appendix S3). We assume that a shipment will always include at least one animal and shift the distributions by 1 to the right, meaning that the corresponding random variable in practice is defined as the number of animals in addition to one animal that is always assumed to be present. Finally, in order to give the parameters associated with these distributions a more straightforward interpretation we parameterized around their mean rather than the standard parameters (i.e. shape and scale or rate for the gamma-Poisson distribution, and two shape parameters for the beta-binomial distribution). With \( i = i_m \) and \( j = j_m \) in this context in order to avoid excessive subscripts, shipment size \( v_m \) is thus given by

\[
v_m = 1 + y,
\]

where \( y \) is the mean size of market-market shipments and, similarly, for the beta-binomial distribution \( \rho \) is defined as the mean proportion of herd size \( h_i \) or \( h_j \) that is shipped. As a consequence of expressing the distributions using their respective means, the parameters \( \psi \) and \( v \) are defined as shape and sample-size, and control the variation around the respective means.

### 2.3.3. Likelihood function

Denoting observations in the shipment data by an asterisk, we define an observed set of individual shipments \( M \) passing between premises \( i \) and \( j \) in quarter \( q \) in the shipment data set \( M^* \) as \( M_{i,j}^* \), and note that the number of shipments is the size of this subset, \( |M_{i,j}^*| = w_{i,j}^* \). The shipment data includes all such observations for which \( |M_{i,j}^*| > 0 \), i.e. at least one shipment is observed, but here we also include all observations for which \( |M_{i,j}^*| = 0 \). Given the current model parameters \( \theta \), the probability of observing exactly \( M_{i,j}^* \) is the joint probability of observing exactly \( w_{i,j}^* \) and the probability of observing exactly the shipment sizes \( v_{i,j}^* \) for each \( m \in M_{i,j}^* )

\[
P(M_{i,j}^* | \theta) = P(w_{i,j}^* | \theta) \prod_{m \in M_{i,j}^*} P(v_m | i,j, \theta) \quad (10)
\]

With the definition of the number of shipments between any given pair of premises as a Poisson distributed random variable, it follows that the first factor of Eq. (10) is given by the Poisson probability mass function,

\[
P(w_{i,j}^* | \theta) = \frac{z_{i,j}^{w_{i,j}^*} e^{-z_{i,j}}}{w_{i,j}^* !}, \quad (11)
\]

where \( w_{i,j}^* \) is given by Eq. (3) and \( z_{i,j} \) is the probability that an interstate shipment that occurred and originated from the state \( s_i \) to \( s_j \) and included in the data set. Since the CVI data consisted of a tenth of all CVI records, \( z_{i,j} \) was set to 0.1 as a baseline value (except for New Jersey, for which it was 0) and then adjusted for each individual state \( s_i \) to reflect the removal of any CVI entries from the state according to Section 2.2.2.

Further, the second factor of Eq. (10) is the product of the individual probabilities of observing the shipment sizes \( v_m \) of each of the observed shipments in subset \( M_{i,j}^* \). These individual probabilities are given by the gamma-Poisson or beta-binomial probability mass functions according to Eq. (9). If there are no shipments between \( i \) and \( j \) in \( q \), the set \( M_{i,j}^* \) contains no elements and the product evaluates to the multiplicative identity (1.0).

Given the above, the full likelihood function for observing the shipment data set \( M^* \), conditional on the model parameters, is simply the joint probability of observing each of the individual subsets \( M_{i,j}^* \),

\[
P(M^* | \theta) = \prod_{i \in P} \prod_{j \in P} \prod_{q \in \{1,2,3,4\}} \prod_{m \in M_{i,j}^*} P(v_m | i,j, \theta) \quad (12)
\]

excluding any subset of observations representing the premises sending to itself \( (j \neq i) \) and any observations of within-state shipments \( (s_i \neq s_j) \). For the beef model \( M^* \) is the 2009 beef CVI interstate shipment data set and \( P = P_{\text{Beef}} \) (as given in Section 2.2.1), while for the dairy model \( M^* \) is the 2009 dairy CVI interstate shipment data set and \( P = P_{\text{Dairy}} \).

#### 2.3.4. Non-standard parameterizations

To facilitate an intuitive interpretation of the kernel parameters, we reparameterized Eq. (8) to be expressed by the parameters \( \kappa_{s, q} \) defined as the distance at which the function has fallen to 50% (\( x_2 = 0.05 \)) of the value when evaluated at \( d = 0 \), and \( x_{s, q} \) defined as the ratio between \( \kappa_{s, q} \) and the distance where the function has fallen to 5% (\( x_2 = 0.05 \)) of the kernel evaluated at \( d = 0 \), yielding

\[
a_{s, q} = \frac{\kappa_{s, q}}{x_{s, q}}, \quad (13)
\]

and

\[
b_{s, q} = \frac{1}{\log x_{s, q}} \left( \log \frac{1}{x_2} - \log \frac{1}{x_{s, q}} \right). \quad (14)
\]

Similarly, we express both of the mixing distributions for the shipment size (Eq. (9)) using non-standard parameters. For the beta distribution the parameters \( \psi_{i,j,q} \) and \( \nu_{i,j,q} \) are related to the standard parameters \( \beta_{i,j,q} \) and \( \gamma_{i,j,q} \) as

\[
\psi_{i,j,q} = \beta_{i,j,q} / (\beta_{i,j,q} + \gamma_{i,j,q}) \quad \text{and} \quad \nu_{i,j,q} = \beta_{i,j,q} + \gamma_{i,j,q} \cdot \quad (12)
\]

In this way \( \psi_{i,j,q} \) is defined as the mean of the beta distribution and translates into the expected proportion of a herd of size \( h \) that is sent. With this definition the parameter \( \nu_{i,j,q} \) is formally the sample size, but can be thought of as controlling the variance of a beta distribution with a given mean \( (\nu_{i,j,q} \beta_{i,j,q} + \gamma_{i,j,q})^{-1} \). The gamma distribution is also expressed in a non-standard way with parameters \( x_2 \) and \( x_{s, q} \), defined as the mean and shape of the distribution, respectively. Using these non-standard formats gives a more concrete meaning to the parameters which makes for easier interpretation and simplifies the choice of prior distributions somewhat. For instance, it is easier to have an idea about the distribution of the expected proportion of a herd that is shipped, i.e. the mean of the beta distribution, rather than the shape parameters of the beta distribution themselves.

### 2.3.5. Priors

In order to simplify the choice of prior distributions, we strove to define the model parameters in ways that allowed meaningful interpretations of their role in the model. Intuitive understanding of the parameters facilitates prior elicitation, and our approach was to, when possible, select a plausible interval and then define the hyperparameters of a suitable prior distribution such that the 2.5th and 97.5th percentiles coincide with the interval. In this way, our defined intervals encompasses 95% of the prior density but still permits more extreme parameter values, should the data strongly contradict our priors. We use the notation \( P_{5, 95} \) and \( P_{95, 5} \) to denote these percentiles. In this section, we detail and motivate our choice of priors, see Table 3 for a summary.

The definition of \( \psi_{i,j,q} \) is a premises type-dependent baseline shipping rate corresponding to what \( u_{i,j,q} \) would be given two immediately adjacent average premises (with respect to herd size, county...
covariates, and state-level random effects). This definition makes it difficult to have an intuition of what constitutes a plausible parameter value, and therefore assigned a vague prior to \( \eta_i \) in the form of a gamma distribution with parameters \( \gamma = 0.001 \).

The parameter \( \zeta_{i,j,q} \in (0, \infty) \) scales the shipment rate \( u_{j,q} \) for a given combination of premises types up or down between quarters. We did not expect the rate to fluctuate by more than a factor ten between quarters and assigned a gamma distributed prior to \( \zeta_{i,j,q} \) with \( P_{2.5\text{th}} = 0.1 \) and \( P_{97.5\text{th}} = 10 \).

The herd-size scaling exponents \( \phi_{i,q}^{\text{in}}, \phi_{i,q}^{\text{out}} \in (-\infty, \infty) \) and \( \phi_i^{\text{in}}, \phi_i^{\text{out}} \in (-\infty, \infty) \) control how the shipment rate between two premises scales with the premises’ herd sizes, as \( \hat{h}_{i,q} \propto \hat{h}_i \phi_{i,q} \). We generally expected larger herd sizes to be associated with higher shipping rates. However, it is also conceivable that large premises send larger shipments less often than smaller premises. Therefore we allowed for both a positive and a negative relationship between herd size and shipping rate by defining the priors for \( \phi_{i,q}^{\text{in}} \) and \( \phi_i^{\text{in}} \) as normal distributions with \( P_{2.5\text{th}} = -0.5 \) and \( P_{97.5\text{th}} = 1.5 \).

Effects of individual cattle industry covariates \( k \) on \( u_{i,j,q} \) are determined by the weights \( r_{i,j,q}^{\text{in}}, r_{i,j,q}^{\text{out}} \in (-\infty, \infty) \). The shipment rate scales exponentially with the product of the individual covariate \( r_{i,q} \) of county \( c \) and its corresponding weight parameter. Thus, a parameter value of zero means no effect of covariate \( r_k \) on the rate, while positive and negative values yield positive and negative relationships between \( r_k \) and the rate. We expected the covariates to have both positive and negative effects on the rate, and therefore assigned a normal distribution with \( P_{2.5\text{th}} = -2.0 \) and \( P_{97.5\text{th}} = 2.0 \) (i.e. centered around zero) as the prior for \( r_{i,j,q}^{\text{in}} \) and \( r_{i,j,q}^{\text{out}} \).

The size of shipments between markets follows a gamma-Poisson distribution with average shipment size \( \xi_{i,q} \in (0, \infty) \) and shape \( \psi_i \in (0, \infty) \). The average lot size marketed through a livestock video auction service between the years 1995 and 2018 was reported to be 121 animals (McCabe et al., 2020). Therefore we defined the prior distribution of \( \xi_{i,q} \) to be lognormal with \( P_{2.5\text{th}} = 12.1 \) and \( P_{97.5\text{th}} = 121.0 \), i.e. one order of magnitude up and down from this average. In order to allow for a wide range of sizes we set the prior of \( \psi_i \) to be a lognormal distribution with \( P_{2.5\text{th}} = 0.1 \) and \( P_{97.5\text{th}} = 100 \).

The mean of the beta distribution in Eq. (9), \( \rho_{i,j,q} \in (0,1) \), is the expected proportion of a farm’s or feedlot’s herd size sent or received. We did not have any prior beliefs for this parameter and a prior distribution uniform on the interval (0,1) was chosen. For the prior parameter \( \psi_{i,q} \in (0, \infty) \), which is the sum of the two shape parameters in the beta distribution and controls the variation around the mean, the prior was set to a lognormal distribution with \( P_{2.5\text{th}} = 2.0 \) and \( P_{97.5\text{th}} = 200 \).

Hierarchical prior structures were used for the parameters of the distance kernel and the state-level random effects \( k_{i,q}^{\text{in}}, k_{i,q}^{\text{out}}, w_{i,q}^{\text{in}}, w_{i,q}^{\text{out}}, a_i^{\text{in}} \) and \( a_i^{\text{out}} \) and the hyperparameters of each of the corresponding prior distribution were estimated as model parameters. Each of these requires a hyperprior to be defined (i.e. a prior for the hyperparameters). With this approach the states with a relative abundance of shipment data were able to inform the parameters of states with less data via the common prior.

The distance kernel is defined with parameters scale, \( s_i \in (0, \infty) \) and shape, \( c_i \in (1, \infty) \) for each state \( s \). The unit of the scale is meters and describe the distance \( d \) at which the kernel has fallen to half of its value at \( d = 0 \). The prior for \( s_i \) was set to a lognormal distribution defined with hyperparameters mean \( \mu^s \), and coefficient of variation \( \xi^s \). Given its definition, we expected the values of \( s_i \) to fall between the shortest and longest plausible shipment distances. The distance between the centroids of the two counties furthest apart in the model is approximately \( 4.6 \times 10^6 \) m which gives an idea of the upper bound. We reasoned that 1000 m is a suitable lower bound as this is close to the scale of the physical size of a premises. Thus, for the hyperprior on \( \mu^s \) we used a lognormal distribution with \( P_{2.5\text{th}} = 10^4 \) m and \( P_{97.5\text{th}} = 4 \times 10^6 \) m. For the coefficient of variation \( \xi^s \) we wanted to allow a span of values large enough to make a wide range of values possible and defined its hyperprior to be lognormal with \( P_{2.5\text{th}} = 0.1 \) and \( P_{97.5\text{th}} = 10 \). For the distance kernel shape parameter, \( \alpha_{i,q}^{\text{in}} \), we strove to construct a prior that allowed for a wide range of function shapes. For this a lognormal prior was again chosen with hyperparameters mean \( \mu^\alpha \), and coefficient of variation \( \xi^\alpha \), each with its own lognormal hyperprior distribution with \( P_{2.5\text{th}} = 1.0 \) and \( P_{97.5\text{th}} = 100 \), and \( P_{2.5\text{th}} = 0.1 \) and \( P_{97.5\text{th}} = 10 \), respectively.

The state-level random effects, \( \omega_{i,q}^{\text{in}}, \omega_{i,q}^{\text{out}} \in (0, \infty) \), were assigned gamma distributed priors defined to have mean \( = 1.0 \) and shape hyperparameters \( a_i^{\text{in}} \) and \( a_i^{\text{out}} \), respectively. This way, the individual state-level parameters can be thought of as samples from a common population where the average state has a random effect of 1.0. States in which the premises have a higher than average propensity to send or receive shipments will have random effects above one, while the opposite is true for lower propensities, all relative to an average state, making the parameter’s interpretation intuitive. As an uninformative hyperprior on the hyperparameters \( a_i^{\text{in}} \) and \( a_i^{\text{out}} \) we used half-Cauchy distributions with scale = 1.0.  2.4. Computation

2.4.1. Sampling of origin and destination premises

The likelihood function Eq. (12) described in the previous section assumes that the herd sizes and types of the premises \( i \) and \( j \) are known, but the CVI dataset used for parameter estimation lacks such information, necessitating a model component which estimates the identities of the origin and destination premises. For this we constructed a Gibbs-sampler that for each shipment \( m \) samples the origin premises \( i_m \) from a set of available premises, conditional on \( \theta \) as well as the current \( j_m \). An equivalent sampler was used to sample the destination premises \( j_m \) conditional on \( i_m \). Each shipment observed in the data contains information about origin and destination county, which determines the population of premises from which to sample \( i_m \) and \( j_m \). For some shipments this space is further restricted to only include certain premises types, as for some shipments information was available that indicated the type of the origin premises or destination premises (or both, see Section 2.2.2).

For a single shipment, \( m \), occurring in quarter \( q_m \), let \( i_m \) and \( j_m \) be the origin and destination premises of the shipment, and let \( c(i_m) \) and \( c(j_m) \) be the counties where the respective premises is located. Further, let \( P_{c(i_m)} \) be the subset of all premises that lie in the origin county \( c(i_m) \) and match the types specified as acceptable for the origin premises of \( m \) according to the CVI data. Equivalently, let \( P_{c(j_m)} \) be the set of premises in the destination county \( c(j_m) \) that match the types specified as acceptable for the destination premises of \( m \). The premises are then sampled according to probabilities proportional to the weights

\[ P(i_m | \theta, j_m, \theta) \propto u_{i_m} P(u_{i_m} | j_m = x, \theta), \quad x \in P_{c(i_m)} \]  

and

\[ P(j_m | \theta, i_m, \theta) \propto u_{j_m} P(u_{j_m} | i_m = x, \theta), \quad x \in P_{c(j_m)} \]
2.4.2. Calculating the joint posterior distribution and assessment of model convergence

We estimated the posterior distributions separately for beef and dairy industries. The beef analysis included the 2009 beef CVI data set and the set of beef premises $P_{beef}$. The dairy analysis included the 2009 dairy CVI data set and the set of dairy premises $P_{dairy}$. For both analyses, the same set of industry covariates were used.

Metropolis–Hastings updates (Hastings, 1970) were used to update most model parameters ($\phi$, $\kappa$, $\pi$, $\omega^{mix}$, $\gamma$, $\rho$, $\nu$, $\chi$ and $\psi$) and all hyperparameters for priors with a hierarchical structure. For the remaining model parameters ($u_i$, $\zeta$, $\omega$), we specified conjugate priors, which permitted Gibbs sampling directly from the respective conditional distributions. One exception was the parameter $\zeta_{i,q}$ for $(i = t_j = farm)$ and $q = 1$, which was fixed at 1.0 to make the model identifiable. All proposals for the Metropolis–Hastings updates were made exclusively from multivariate normal distributions with adaptive scaling as described in Garthwaite et al. (2016). Scalars $\rho_{i,j,q}$, $\nu_{i,j,q}$, $\chi_{i,q}$ and $\psi_{i,q}$ for all values of $t_i$, $t_j$ and $q$ were updated univariately, whereas some form of joint update was implemented for the remaining parameters. For $\phi$ and $\gamma$, each vector $\phi_i$ and $\gamma_i$ associated with each quarter was updated separately, with the scalars $\phi_{i,t}$ and $\phi_{i,q,j}$ for all types $i$, and $\gamma_q$ consisting of the scalars $\gamma_{q,t}$ and $\gamma_{q,j}$ for all covariates $k$.

It would have been possible to sample $\omega^{mix}$ from its conditional distribution. Yet, because the posterior was highly correlated with the posterior distributions of the kernel parameters $\kappa$ and $\pi$, we improved mixing by including $\omega^{mix}$ in a joint proposal together with $\kappa$ and $\pi$ for each combination of origin state $s$ and quarter $t$. The hyperparameters $\mu^s$ and $\kappa^s$ were updated jointly, as were $\mu$ and $\kappa$, while $\omega^{mix}$ and $\omega^{mix}$ were updated separately. The parameters $\kappa$, $\pi$, $\omega^{mix}$, $\nu$, $\chi$, $\psi$, $\mu$, $\kappa$, and $\omega^s$ were proposed on the log scale, while $\rho$ was proposed on the logit scale.

For beef, ten MCMC chains were run, each for 400,000 iterations. This was close to the upper limit of the number of iterations that could be achieved within the seven day job limit imposed by the computing cluster used. The first 50,000 samples were discarded as burn-in, and the remaining 350,000 samples were thinned by selecting every 100th sample. Thus, with ten chains the final beef posterior distribution consisted of 35,000 samples. Dairy analyses were substantially faster, and we ran ten chains for 700,000 iterations, out of which the first 50,000 were discarded. The remaining 650,000 thinned to 6500 samples and the ten posteriors were combined giving a final sample size of 65,000.

To evaluate computational performance, we estimated the effective number of independent samples, $n_{eff}$ (Gelman et al., 2004). One purpose of the study was to simulate 1000 synthetic networks, and the specific number of iterations and parallel MCMC chains to run was chosen in order to satisfy at least $n_{eff} = 1,000$ for all parameters. In addition, we also monitored the potential scale reduction factor $\hat{R}$, which gives an indication of whether the parameter estimates have converged, with values close to 1.0 indicating convergence (Gelman et al., 2004).

The entire statistical model was implemented in C++, compiled with GCC 7.3.0 and parallelized using OpenMP. The parameter estimation steps and the network simulations were performed on the 32-core compute nodes of the Tetralith High Performance Computing cluster at the National Supercomputer Center (NSC) in Linköping, Sweden.

2.5. Simulation of networks

Using the estimated parameters, we simulated 1000 complete beef and dairy networks, each based on a single joint random sample from the respective posterior distribution. The simulation procedure consisted of calculating the shipment rate $n_{ij}$ for every single pair of premises $i$ and $j$ in the relevant premises data and sampling the number of shipments from $i$ to $j$ in quarter $q$ as $w_{ij,q} \sim Pois(n_{ij,q})$. For each shipment, a size was then sampled according to Eq. (9).

When calculating $u_{ij,q}$ for two premises within the same county $c$, the estimated distance between the two premises $(d_{ij,c}$ used in the distance kernel function, Eq. (8)), was set to be the average distance between random coordinate pairs inside a square with the same area as the county, i.e. $d_{ij,c} = 0.5214\sqrt{A_c}$ where $A_c$ is the area of county $c$ (Bäsel, 2021).

2.5.1. Model validation

To validate the simulated networks against the training data, the networks were down-sampled to mimic the way the CVI-data were collected. Thus, we removed all within-state shipments and (randomly) a proportion $1 - z$, of the remaining shipments for each origin state $s$. The simulated networks were then compared to the CVI data at the annual level using several network metrics. For the purpose of this network analysis, a node is defined as a county, and shipment connections as directed edges. The edges are weighted by both number of shipments and number of animals. The following network level metrics were considered: number of active nodes, number of edges per node, reciprocity, transitivity and fragmentation. Reciprocity is defined as the proportion of edge pairs with an edge between them that also have an edge in the opposite direction (Wasserman and Faust, 1994). Transitivity, or clustering coefficient, is the proportion of all possible node triplets that are fully connected or, in other words, the extent to which nodes’ neighbors are also connected to each other (Wasserman and Faust, 1994). Fragmentation measures to what extent the network is divided into separate components and is defined as the proportion of nodes between which no path exists (Borgatti, 2006). The following metrics at the node level were considered: in-degree, out-degree, and weighted betweenness centrality (Freeman, 1977). Betweenness centrality is defined as the proportion of shortest paths between all node pairs that pass through the node in question. In the case of weighted betweenness centrality the weights are treated as costs (or distances) which increase the effective path lengths, so for this purpose we used the reciprocals of number of animals and number of shipments as weights. The analyzed network metrics are of interest for e.g. disease spread analyses, and the network analysis was used to enable posterior predictive checks as part of the model validation to ensure relevant features were recapitulated. The network analysis was done using the python interface to the iGraph C library (Csardi and Nepusz, 2006).

We also performed out-of-sample validation against two additional CVI data sets from the years 2010 and 2011. These data encompassed only a subset of origin states (CA, IA, MN, NC, NY, TN, TX and WI), consisted of a 30% random sample of interstate shipments, and were not separated into beef and dairy. In order to enable comparison against these additional data sets, the simulated networks were down-sampled to correspond to this sampling scheme. Two comparisons were made against the 2010 and 2011 data. First, the shipment size distribution of the simulated networks was validated against the shipment size distribution of the data sets. For this comparison, beef and dairy networks were pooled together by concatenating the beef network that was simulated first with the first dairy network, and the second beef networks were pooled together by concatenating the beef network that was simulated first with the second dairy network, and so on. This was necessary to make the comparison possible, as the 2010 and 2011 data sets made no distinction between beef and dairy shipments. Second, for every origin state in the subset, we calculated the correlation between the simulated number of shipments to each other state, and the number of shipments to each other state in each data set. This was done for every simulated network in order to obtain a distribution of correlations to the two data sets for each origin state.

3. Results

3.1. Posterior marginal distributions of estimated parameters

The model parameters were defined with interpretability in mind, and their estimated posterior distributions reveal important information about the system. Due to the large number of model parameters,
we only present a summary of the marginal distribution of a subset of parameters here. A complete set of summaries of the marginal posterior distributions can be found in the supplemental information. Additionally, for the baseline shipment rate, we report the quantity $\hat{h}_{i,t|q} = \hat{n}_{i,t} \hat{s}_{i, j|q}$. This metric is the baseline shipping rate by premises types and quarter and is likely of greater interest than any of the two model parameters themselves.

3.1.1. Beef

For the parameters defined at the premises level, most showed clear effects of premises type and time of the year on the shipment rate $u_{i,t|q}$. The herd-size scaling parameters ($\phi^{\text{out}}_t$ and $\phi^{\text{in}}_t$, Fig. 1, panels A and B) indicated a linear relationship between feedlot herd size and shipment rate when acting as the receiver of shipments, with $\phi^{\text{in}}_t\text{_feedlot}$ close to 1.0, but a weaker relationship when a feedlot was the origin premises. For farms, the relationships were close to the opposite, with high scaling factors estimated for origin premises and low values for receiving premises. For markets, the effect was intermediate in comparison to the two other premises types for both directions. The herd-size scaling parameters also displayed one of the most dramatic effects of quarter among the parameters at the premises level (Fig. 1), although there was no clear interaction with premises type. For instance, while feedlots exhibited a negative relationship between shipment rate and herd size when sending shipments in quarter three, farms showed the highest positive effect the same quarter and nearly no effect of herd size in quarter one and two. Markets had similar seasonal patterns when receiving and sending shipments, with the highest scaling factor in quarter four and the lowest in quarter two, although the differences were more pronounced when receiving shipments.

The baseline quarterly shipping rate $\hat{h}_{i,t|q}$ was also strongly influenced by premises types (Fig. 2A). This was almost exclusively due to variation in the parameter $\hat{n}_{i,t}$ (Figure S2). For $\hat{\zeta}_{i,t|q}$ — the other component of $\hat{h}_{i,t|q}$ — an effect of premises type was only seen for the farm-to-farm specific parameter, while the parameters for the other premises type combinations showed very little variation over the year (Figure S2).

The largest shipment sizes in relation to herd size ($\rho_t$, Fig. 2B) were estimated for feedlot–feedlot shipments and for farm–feedlot, while the lowest estimates were found for feedlot–farm shipments. Feedlot–farm shipment sizes varied substantially across the year with the estimates for quarters two and three being roughly an order of magnitude higher than for the other two quarters. In contrast, very little variation across the year was found for the estimate for shipments from market to feedlot. The proportional shipment-size parameter for the other combinations of premises types fell somewhere in between these two extremes in terms of variation across the year. The parameter $\nu_t$, which controls the variation of shipment size around $\rho_t$, showed a clear dependence on both premises type combinations, and quarters (Figure S3). The average market–market shipment size ($\phi_{ij}$) was estimated to be lowest in the first quarter of the year and to increase successively during the following quarters (Fig. 2C). The shape parameter of the Gamma-Poisson distribution for market-to-market shipments ($\omega_{i,j}$) was estimated to be similar for quarters two through four, but substantially lower in quarter one (Figure S3).

The covariates related to the cattle industry also showed clear effects on the beef shipment rates, with the scaling factors $\phi^{\text{out}}_t$ and $\phi^{\text{in}}_t$ estimated to have the bulk of the posterior densities $\neq 0$ (Figure S4). The only exception was found for the number of sales measured in head, which showed only a slightly positive effect on $u_{i,t|q}$. The number of operations with sales had a negative effect on the shipment rate for both receiving and sending premises, while the total number of cattle in the county had a positive effect for both receiving and sending premises. The slaughter connectivity measure for the origin county showed a negative effect on $u_{i,t|q}$, but a positive effect in the receiving county.

Parameters defined on the state-level ($\kappa_{i,q}, \pi_{i,q}, \omega_{i,q}^{\text{out}}$ and $\omega_{i,q}^{\text{in}}$) showed spatial variation, and, to a lesser extent, temporal heterogeneity. These differences were less pronounced in the kernel parameters $\kappa_{i,q}$ and $\pi_{i,q}$ (Figures S5 and S6) compared to the state-level random effects $\omega_{i,q}^{\text{out}}$ and $\omega_{i,q}^{\text{in}}$ (Figures S7 and S8).

3.1.2. Dairy

For dairy shipments, the herd-size scaling parameters for farms showed close to linear relationship with size for both the origin and destination herd in all quarters (Fig. 3A and B). For feedlots the same was true for the effect of herd size on incoming shipments as well as for outgoing shipments in quarter 3. In quarter 2 and 4, the origin herd size-scaling for feedlots was still positive if not as high, but for quarter 1 feedlots showed a relationship between herd size and outgoing shipment rate. For markets the scaling exponent for origin premises was close to zero during all quarters, while for receiving premises it was close to one for quarter four, close to zero for quarters one and three, and slightly below zero for quarter two (Fig. 3A and B). Similar to beef, the baseline quarterly shipment rate, $\hat{h}_{i,t|q}$, showed relatively little effect of quarter but a clear effect of premises type with farm-to-farm baseline rate being the lowest and market-to-market rate being the highest (Fig. 4A and Figure S9). The parameters for shipment size as a proportion of herd size, $\rho_{i,t|q}$, showed little effect of either premises type or quarter in the interaction between farms and markets, but plenty of both seasonal and type-related variation for interactions involving feedlots (Fig. 4B). The parameter controlling the variation around this proportion, $\nu_{i,t|q}$, also showed substantial variation among the interactions of different premises types, as well as seasonal variation (Figure S10). The average market–market shipment size parameter, $\chi$, stayed relatively stable over the year, with only quarter two showing a slightly higher estimate than other quarters (Fig. 4C).

When considering dairy state-level parameters, the random effect on incoming shipments, $\omega_{i,q}^{\text{in}}$ showed a much heterogeneity among the destination states, as well as some seasonal variation. The kernel parameters and the random effect on outgoing shipments ($\kappa_{i,q} \pi_{i,q}$ and $\omega_{i,q}^{\text{out}}$), however, showed substantially less variation among most states and quarters (Figures S11–S14).

Out of the industry covariate scaling parameters, the slaughter connectivity (SC) measure showed a clear positive effect on $u_{i,q}$ for dairy shipments for the origin premises, while the other industry covariates showed only modest effects on the origin side (Figure S15A). For the destination side, only the number of operations with sales (OS) showed a negative effect, while the credibility intervals for the other covariates all overlapped zero.

Fig. 1. Beef 95% credibility intervals of the herd size scaling exponents $\phi^{\text{out}}_{t}$ (A) and $\phi^{\text{in}}_{t}$ (B) marginal posterior estimates. Dashed lines indicate 1.0 (linear scaling with herd size) and 0.0 (no scaling with herd size).
beef networks was state shipments. The total annual median number of shipments in the simulated 1000 complete beef and dairy networks, including within-

3.2. Simulated networks

3.3. Within-sample model validation
transitivity was lower. For dairy, both reciprocity and transitivity were slightly higher for the simulated networks than for the data. Reciprocity in beef and the number of edges per node and fragmentation for dairy were very close to those of the data.

We analyzed the spatial distribution of the node-level measures in-degree, out-degree and betweenness centrality of the simulated networks and the 2009 CVI data set. As the addition of shipment size is the most important novel contribution of the model, we prioritize showing the results for the flow of animals, i.e. networks with links weighted by number of animals, and present metrics for unweighted links and weighted by number of shipments in the supplementary information.

For both beef (Fig. 9) and dairy (Fig. 10) the spatial distributions of animal flow generally showed very high similarity with those based on the data, though some disparity can be observed. The most obvious difference can be seen in the dairy maps, where contrast between counties is lost somewhat in the simulated networks, making the effect of the overestimation of active nodes (i.e. counties) for dairy apparent.

For the number of shipments sent, the patterns are largely the same (Figures S20 and S21). In the supplemental information, we also report the spatial distribution of in- and out-degree in terms of number of shipments at the state level in order to enable comparison to previous versions of USAMM (Figure S22).

### 3.4. Out-of-sample model validation

To ensure that the model is useful outside of the context of replicating the training data we also performed two comparisons with two sets of CVI data from a subset of states for the years 2010 and 2011. Similarly to the within-sample validation, for this comparison the simulated networks were down-sampled to match the way the data sets were sampled. In these two data sets, information on whether the shipment was beef or dairy does not exist and therefore we compare the combined simulated beef and dairy networks to the data. Similar to the within-sample validation, the comparison to additional data
Fig. 6. Median number of shipments (left) and shipped animals (right) per quarter in the simulated complete dairy networks (n = 1000) divided into the numbers from premises types (rows) to premises types (columns). Each separate triangle represents the median for one quarter of the year.

Fig. 7. Frequency distributions of the total number of shipments (top) and animals shipped (bottom) in the down-sampled simulated beef (left column, n = 1000) and dairy (right column, n = 1000) networks. The corresponding value of the 2009 CVI training data is indicated by the black vertical line.

demonstrated that predictions captured the distribution of shipment sizes of the observed networks (Fig. 8, panels C and D).

For most states, the simulated networks showed a high correlation with the 2010 and 2011 data sets in the number of shipments to different destination states (Fig. 11). A clear exception was seen for North Carolina, where the median correlation across the networks was below 0.4. However, this substantial drop in correlation compared to other states is observed for the correlation of the 2009 CVI data with the other data sets as well.

Table 4

<table>
<thead>
<tr>
<th>Yearly networks</th>
<th>Network metric</th>
<th>CVI data set</th>
<th>Simulated networks, median (95% cred. int.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef</td>
<td>Number of nodes</td>
<td>2322</td>
<td>2487 (2454, 2521)</td>
</tr>
<tr>
<td></td>
<td>Edges per node</td>
<td>4.79</td>
<td>5.77 (5.67, 5.88)</td>
</tr>
<tr>
<td></td>
<td>Reciprocity</td>
<td>0.052</td>
<td>0.054 (0.049, 0.059)</td>
</tr>
<tr>
<td></td>
<td>Transitivity</td>
<td>0.049</td>
<td>0.057 (0.055, 0.060)</td>
</tr>
<tr>
<td></td>
<td>Fragmentation</td>
<td>0.68</td>
<td>0.57 (0.55, 0.60)</td>
</tr>
<tr>
<td>Dairy</td>
<td>Number of nodes</td>
<td>882</td>
<td>1264 (1218, 1304)</td>
</tr>
<tr>
<td></td>
<td>Edges per node</td>
<td>1.96</td>
<td>1.97 (1.90, 2.04)</td>
</tr>
<tr>
<td></td>
<td>Reciprocity</td>
<td>0.092</td>
<td>0.049 (0.038, 0.061)</td>
</tr>
<tr>
<td></td>
<td>Transitivity</td>
<td>0.036</td>
<td>0.025 (0.020, 0.030)</td>
</tr>
<tr>
<td></td>
<td>Fragmentation</td>
<td>0.91</td>
<td>0.89 (0.87, 0.91)</td>
</tr>
</tbody>
</table>

To assess the relevance of the shipment data sets from 2009, 2010 and 2011 for analysis of the present day U.S. cattle industry we visualized trends in the number of cattle operations, number of cattle, and number of cattle on incoming shipments at the state level from 1980 to 2022 according to the NASS survey (Figure S23). No clear trend could be seen in the number of incoming animals, while some states showed a very slight downwards trend in the number of head. The most evident trend was seen for the number of cattle operations which decreased for most states, although the largest state in terms of the number of cattle premises, Texas, was an exception, showing no apparent change over the period. However, no data was available for the number of cattle operations after 2007, making conclusions difficult to draw for the period after which the shipment data was collected.

3.5. Model convergence and computation

The MCMC showed good convergence for both beef and dairy, with the highest values for $\hat{R}$ at 1.0028 and 1.0040 for any single parameter of the respective parameterizations. For most model parameters in either of the two parameterizations, $\hat{n}_{eff}$ was well over 1000 (Figure S24). However, autocorrelation was high for certain parameters —for
Computers and Electronics in Agriculture 203 (2022) 107483

S. Sellman et al.

Fig. 8. Distribution of shipment size in the simulated networks (whiskers and x) compared to the 2009 CVI training data (bars; A and B) and CVI data for the additional years 2010 and 2011 (bars, C and D). The whiskers show the 95% credible interval together with the median of the observed values (x) in each bin across the 1000 simulated beef (A and C) and dairy (B and D) networks, respectively. The simulated networks were down-sampled to match the sampling procedure of the respective data-set.

beef the worst performance in terms of model mixing was shown by the parameter $\zeta_{i,j,q}$ for $t_i = \text{farm}$, $t_j = \text{farm}$ and $q = 4$ for which $n_{\text{eff}} = 1,718$. For dairy the worst performance was seen for $\rho_{i,j,q}$ for $t_i = \text{Feedl}$, $t_j = \text{farm}$ and $q = 3$ for which $n_{\text{eff}}$ was 1236. Considering that all parameters reached an independent effective sample size of at least 1000, we deemed the final joint posterior distributions to be suitable for the purpose of simulating 1000 representative synthetic beef and dairy networks. Executing a chain of 400,000 iterations took approximately 6 days and 7 h on a 32-core compute node for beef data. For the dairy analysis, which used considerably less data, the equivalent run time for 700,000 iterations was 6 days and 14 h.

4. Discussion

Here, we have described the development of a new version of the U.S. Animal Movement Model (USAMM) with the capability to predict the size and spatial distribution of cattle shipments. Using USAMM, we simulated 1000 beef and dairy networks, and validation showed that this new version of USAMM was able to capture most network properties along with large-scale spatial patterns and shipment sizes seen in both the 2009 CVI training data sets and two additional smaller data sets from subsequent years.

A model capable of predicting shipment size is essential, particularly to support modeling of infectious livestock disease in the U.S., where shipment size information is not available. The value of an informed model for shipment size varies with the focal disease. For instance, when modeling FMD, Tsao et al. (2020) made the assumption that any shipments from infected herds, regardless of size, always transmit the infection to the receiving herd, while Garner and Beckett (2005) implicitly included the risk of transmission via shipments together with other transmission pathways in a single function of distance which was fit to outbreak data. However, for other infectious diseases, such assumptions are not always realistic, and adding information that allows weighing risks of exposure based on herd-level and shipment-level prevalence would make allow for increased model realism. For instance, the comparably low herd-level prevalence and within-herd transmission of e.g. bTB means that whether or not just one or a few infected animals are part of a shipment can play a major role in the risk of spread. Therefore, keeping track of individual infected animals and how they move between herds becomes important. Another example is bovine viral diarrhea (BVD), with which individual animals may become persistently infected. Such an individual has a high potential of infecting another herd via movement (Fray et al., 2000), something that a model of between-herd spread of BVD must account for. However, modeling the movement of individual animals between herds in the U.S. requires information about shipment sizes, which the new version of USAMM provides.

To capture important patterns of the shipping network with USAMM, we incorporated several relevant features of the origin and destination herds. It is reasonable to expect herd size to play a role in shipment size, as larger herds will likely be involved in moving more animals. From previous studies, it is also clear that premises of different types show different shipment patterns (Ortiz-Pelaez et al., 2006; Robinson and Christley, 2007; Lindström et al., 2011; Ribbens et al., 2009). Therefore, we included herd size and type of premises as explanatory variables for shipment size. From the posterior distributions of the parameters $\rho$ and $\nu$, which controls the expected shipment size, it is clear that premises type heavily influences both beef and dairy shipment size (Fig. 2B, Fig. 4B, and A-panels in Figures S3 and S10).
In addition to the parameters modeling shipment size, a number of model features affecting the shipment rate between herds were also defined at the premises level. Out of these, only $\eta$ (for both beef and dairy), the parameter controlling the premises type specific temporal variation in shipment rate, did not show a substantial effect of premises type. This indicates that, to the degree with which there is temporal variation in shipment rate, this is similar across production types. The substantial disparity in other premises type specific parameters reveals the importance of including premises type when modeling such systems. Previous studies have already indicated the importance of markets in cattle shipment networks (Robinson and Christley, 2007; Vidondo and Voelkl, 2018), and in the pig industry, the type of a premises strongly influence its contact patterns (Lindström et al., 2011). This heterogeneity is also in line with expectations about how the U.S. beef industry works where, in broad terms, animals are reared in locations separate from where the animals are fattened in feedlots to which they are sent via markets or sometimes backgrounding herds (Shields and Mathews, 2003). With high degree of specialization in premises, differences are expected among these model parameters, something which USAMM clearly captures.

Our model reveals heterogeneous contact structures and assortative mixing patterns in the shipment rates between premises that depend on the types of the premises involved. Considering that shipment size is also very dependent on premises types (compare for instance relative size for farm–farm and farm–feedlot shipments in Fig. 2) it is apparent that a model where the number of shipments and shipment size are not estimated jointly will fail to capture important interaction effects between the two components. However, as we have shown here, a model that captures these interaction effects is complex, with run-times measured in days on highly parallel and efficient computational systems. Such complexity is often too unwieldy to include as a separate component in modeling studies where the research focus is elsewhere.
This has forced modelers to make compromises at the expense of exactness. Kao et al. (2018), for instance, constructed a U.S. bTB model which included between-herd transmission via shipments modeled by an earlier version of USAMM. Without shipment size information, the authors had to infer shipment size from separate sources of premises-level averages of purchases of cattle and number of shipments received. Although this is not a very robust approach, expecting researchers to develop their own detailed shipment model when that is not the focus of the question is setting the bar high. Therefore such simplification may be the only feasible alternative in the absence of shipment size data. However, as a model dedicated to providing a complete picture of the U.S. cattle shipment networks, we hope that the updated version of USAMM presented here provides an alternative to compromise in such situations.

The validation of shipment size against both the training data and the two additional data sets from 2010 and 2011 showed that USAMM does well in terms of replicating the inter-state shipment size distributions. However, one potential weakness of the model is that the within-state shipment size predictions are based purely on parameters estimated from interstate shipment data. The reason was lack of suitable intra-state data, but nevertheless, an implicit assumption that intra- and interstate shipments are equivalent in terms of shipment size lies underneath it. Without additional data, it is of course difficult to validate the intra-state shipment size predictions made by USAMM. One piece of evidence in favor of the model is that there is no apparent relationship between shipment size and distance between the origin and destination counties in the 2009 CVI data (Pearson’s correlation 0.06 and 0.18 for beef and dairy respectively, Figure S25). However, this does not exclude that some other unobserved process is at work at exclusively within states that influences intrastate shipment size.

Compared to previous versions of USAMM, this new addition adds several additional features and a significant amount of complexity, both
in terms of additional parameters and computational load. Although the increased complexity is necessary in order to achieve the goal of predicting shipment sizes, it is conceivable that such an increase could detract from the model's previous capability to predict shipment patterns. However, compared to the most recently published version of USAMM (Brommesson et al., 2021), this new version does at least as well at the state level in terms of in- and out-degree (Figure S22). Additionally, the analysis of the correlation in the pattern of number of state-to-state shipments showed a result that was very much in line with what was seen in Brommesson et al. (2021). This indicates that not only are the degree distributions preserved from earlier studies, but the way the shipments from each origin state are distributed among the receiving states is also similar, showing that the model that includes small scale prediction, including the size of individual shipments, does not perform worse at the coarse scale than previous versions.

Spatially, the model convincingly recreates the general geographical patterns seen in the training data with regard to in- and out-degree and betweenness centrality (Figs. 9 and 10). However, in some areas, the degree measures of the simulated networks show a slight tendency of dispersing shipped animals over more counties compared to the data. This is apparent as lower contrast in figures (Figs. 9 and 10, e.g. beef out-degree in Oklahoma, and dairy in-degree in Arizona). This indicates that there may still be a county-level property not captured by our industry covariates that differentiates counties in terms of how likely their premises are to send and receive shipments.

Although the validation step showed satisfactory results, the entire out-of-sample validation is hampered by the lack of detailed shipment data to compare the model predictions against. In particular the two data sets that were used for the out-of-sample validation contained no information of whether shipments were for beef or dairy. This required the simulated beef and dairy networks to be pooled into combined networks which reduces the level of detail in the comparison. Moreover, no information about premises type is present in these two data sets, precluding an analysis of how well our model captures premises type specific shipment behavior. Likewise, the lack of within-state shipment size information prevents validation of the model's predictions of such shipments.

The 2009 CVI data set constitute the only collection of cattle shipments for all 48 conterminous U.S. states and, given its age, it is possible that it no longer reflects the current state of the cattle industry accurately. It has long been recognized that there is a shift towards larger and more specialized agricultural operations in the U.S. McDonald and Coffman (1980), Baltensperger (1993) and MacDonald and McBride (2009), and this shift may have had some impact on the cattle trade network since the collection of the data sets used here. However, quantifying any such change with a reasonable degree of certainty is a challenge due to the scarcity of accessible data. Our analysis of the historical state-level survey data available via NASS generally confirmed this trend for the number of cattle operations (Figure S23), but it is difficult to assess how this affects the relevance of the shipment data for predicting the present day network as the time series for this variable ended in 2007. For the number of incoming animals to the state, and the state-level cattle inventory, no clear trends are discernible over the period since the shipment data was collected, and analysis of the three years of CVI data suggests that the overall spatial pattern of livestock shipments are mostly stable over this time, mainly due to fixed infrastructure related to the cattle feedlot–slaughter chain Gorsich et al. (2016). Given that only the number of operations seem to be declining slowly, we believe that the predictions made by USAMM are relevant for the present day situation, and it is likely that the observed change in the demography has little effect on the general structure of the networks. However, through computer simulations Meadows et al. (2018) were able to show that higher concentration of animals on fewer premises is linked to more severe outbreaks of FMD in the U.S., and it is not unreasonable to expect that a similar increase in the density of shipments between fewer premises could have a similar effect on other infections. We acknowledge that the shipment data may be considered out-of-date, and that a more recent collection of CVIs would be very useful—not only for USAMM, but also for the broader understanding of temporal trends it could provide though studies similar to that of Gorsich et al. (2016).

5. Conclusion

The set of interstate certificates of veterinary inspections (CVI) from 2009 collected by Buhrerkepmpe et al. (2013) continues to be the most comprehensive source of information with complete national cover for U.S live cattle movement, but it lacks the level of detail necessary for epidemiological modeling. This prompted us to develop an advanced version of USAMM for cattle capable of taking into account herd-level characteristics in order to scale up the CVI data to full networks of inter- and intrastate shipments across the entire conterminous U.S. at the resolution of individual premises. This increase in resolution compared to previous models enabled us to leverage additional, previously unused, information about shipment size present in the CVI data set to make predictions about the number of animals shipped in addition to the spatial structure of the cattle trade network. Additionally, the finer scale of the model has enabled us to expose the relationships of premises size and premises type with shipment patterns, giving...
previously unobtainable insights about the U.S. live cattle trade network. The lack of models that predict the sizes of cattle shipments has hampered epidemiological models of slow moving diseases like bTB in particular, and has limited the usefulness of simulated U.S. cattle networks to situations where shipment size is of little importance. Consequently, the development presented here has opened up the possibility of using USAMM in a broader spectrum of applications. To help further research involving the U.S. cattle trade patterns, we make the 1000 beef and dairy networks publicly available in their raw form at http://dx.doi.org/10.25675/10217/234115, and additionally as an interactive online explorative tool at https://usamm-gen-net.shinyapps.io/usamm-gen-net/.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Funding

This work is supported by funding provided by the U.S. Department of Homeland Security Science and Technology Directorate under contract number D15PC00278.

This work was also supported by USDA Cooperative Agreement: AP17VSCEAH0001012.

SS, LBJ, CW and TL were supported by D15PC00278. ROB was supported by AP17VSCEAH0001012. The findings and conclusions in this publication have not been formally disseminated by the US Department of Agriculture and should not be construed to represent any agency determination or policy. The analyses, views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the regulatory opinions, official policies, either expressed or implied, of the U.S. Department of Homeland Security.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compag.2022.107483.

References


Robinson, S.E., Christley, R.M., 2007. Exploring the role of auction markets in cattle
Okafor, Chika C., Grooms, Daniel L., Bruning-Fann, Colleen S., Averill, James J., Ka-
O’Hare, A., Orton, R.J., Bessell, P.R., Kao, R.R., 2014. Estimating epidemiological
Meadows, Amanda J., Mundt, Christopher C., Keeling, Matt J., Tildesley, Michael J.,
Palisson, A., Courcoul, A., Durand, B., 2016. Role of cattle movements in bovine
tuberculosis spread in France between 2005 and 2014. Prevent. Vet. Med. (ISSN: 0167-
7995246979813&doi=10.1111%2fj.1865-1682.2011.01187.x&partnerID=40&md5=
Robinson, S.E., Christley, R.M., 2007. Exploring the role of auction markets in cattle