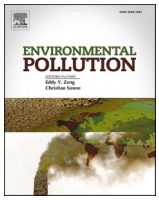
Environmental Pollution 346 (2024) 123590

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol



Residential exposure to microbial emissions from livestock farms: Implementation and evaluation of land use regression and random forest spatial models☆

Beatrice Cornu Hewitt *, Lidwien A.M. Smit, Warner van Kersen, Inge M. Wouters, Dick J.J. Heederik, Jules Kerckhoffs, Gerard Hoek, Myrna M.T. de Rooij

Institute for Risk Assessment Sciences (IRAS), Division of Environmental Epidemiology, Utrecht University, Utrecht, the Netherlands

ARTICLE INFO

Handling Editor: Prof. Dr. Klaus Kümmerer

Keywords:
Air pollution
Microbial emissions
Livestock farming
Spatial modelling
Residential exposure
Antimicrobial resistance

ABSTRACT

Adverse health effects have been linked with exposure to livestock farms, likely due to airborne microbial agents. Accurate exposure assessment is crucial in epidemiological studies, however limited studies have modelled bioaerosols. This study used measured concentrations in air of livestock commensals (*Escherichia coli (E. coli)* and *Staphylococcus* species (spp.)), and antimicrobial resistance genes (*tetW* and *mecA*) at 61 residential sites in a livestock-dense region in the Netherlands. For each microbial agent, land use regression (LUR) and random forest (RF) models were developed using Geographic Information System (GIS)-derived livestock-related characteristics as predictors. The mean and standard deviation of annual average concentrations (gene copies/m^3) of *E. coli*, *Staphylococcus* spp., *tetW* and *mecA* were as follows: 38.9 (±1.98), 2574 (±3.29), 20991 (±2.11), and 15.9 (±2.58). Validated through 10-fold cross-validation (CV), the models moderately explained spatial variation of all microbial agents. The best performing model per agent explained respectively 38.4%, 20.9%, 33.3% and 27.4% of the spatial variation of *E. coli*, *Staphylococcus* spp., *tetW* and *mecA*. RF models had somewhat better performance than LUR models. Livestock predictors related to poultry and pig farms dominated all models. To conclude, the models developed enable enhanced estimates of airborne livestock-related microbial exposure in future epidemiological studies. Consequently, this will provide valuable insights into the public health implications of exposure to specific microbial agents.

1. Introduction

A growing body of epidemiological research demonstrates links between residential proximity to livestock farms and adverse health effects (Radon et al., 2007; Schulze et al., 2011; Borlée et al., 2015; Mirabelli et al., 2006; Pavilonis et al., 2013; Schinasi et al., 2011; Sigurdarson and Kline, 2006). Air pollution originating from these farms is a likely culprit, highlighting the need for reliable predictive models to evaluate airborne concentrations. Historically, environmental impact studies have primarily focused on ammonia, dust and methane emissions from livestock farms and have thus developed models for these emissions (Pohl et al., 2017; de Rooij et al., 2018; Chai et al., 2014; de Vries et al., 2023). However, the biological components of livestock farm emissions have received limited attention despite their substantial emission quantities and potential significance for public health. Limited research exists concerning the evaluation and modelling of livestock-related microbial air pollution.

Livestock farms can emit various biological agents, including bacteria and antimicrobial resistant bacteria harbouring antimicrobial resistance genes (ARGs) (Davis et al., 2018; de Rooij et al., 2019a; Franceschini et al., 2019; Gao et al., 2023; Gibbs et al., 2006), therefore these are good indicators of livestock-related microbial emissions that can be used for modelling. The bacteria *Escherichia coli (E. coli)* and *Staphylococcus* species (spp.) are both well-known livestock commensals, and elevated concentrations have been identified in the air in close proximity to farms (de Rooij et al., 2019a). Moreover, ARGs are commonly detected in the air around livestock farms, with higher levels observed downwind compared to upwind (Gibbs et al., 2006). The ARGs *tetW* and *mecA* confer resistance to antimicrobial classes that are widely used in livestock (tetracycline and beta-lactam antibiotics, respectively)

☆ This paper has been recommended for acceptance by Prof. Dr. Klaus Kümmerer.

* Corresponding author. Institute for Risk Assessment Sciences (IRAS), P.O. Box 80178, 3508 TD Utrecht, the Netherlands.
E-mail address: b.cornuhewitt@uu.nl (B. Cornu Hewitt).

https://doi.org/10.1016/j.envpol.2024.123590
Received 29 September 2023; Received in revised form 10 January 2024; Accepted 14 February 2024
Available online 20 February 2024

(Robles-Jimenez et al., 2021). In the Netherlands, the substantial use of veterinary antibiotics, particularly tetracycline and broad-spectrum penicillins (including beta-lactams), led to policy interventions in 2007 (Mevius and Heederik, 2014). Despite a significant decrease since then, antimicrobial usage on livestock farms continues to exceed human antimicrobial use (Jaarverslag, 2015). As reliable and specific indicators of livestock-related microbial emissions, these commensal bacteria and ARGs were selected as candidates for modelling in this study.

Although spatial models have been developed for various chemical air pollutants in the past, only recently have researchers started to explore the potential of models for predicting concentrations of bio-aerosols (de Rooij et al., 2018; Hjort et al., 2016). Modelling of microbial emissions poses distinct challenges. Microbial emissions, being living organisms subject to biological processes, are likely to exhibit more spatial heterogeneity compared to that of chemical emissions. Given this likely increased spatial heterogeneity, accurately capturing microbial emission patterns becomes challenging. Therefore, it is important to investigate the use of empirical models for livestock-specific microbial emissions to assess the effectiveness of such models for these types of emissions and to identify the most suitable modelling approach. Additionally, this study contributes to the broader understanding of employing spatial models for biological agents, thereby enhancing exposure assessment and facilitating the implementation of effective strategies for public health management.

Historically, estimating residential exposure to livestock-related emissions relied on simple proxies like distance to the nearest farm. While proxies provide reasonable estimations of exposure, they lack the ability to identify specific agents relevant to particular health outcomes and they do not consider emission patterns. In a previous study, the spatial distribution of endotoxin, a component of Gram-negative bacterial cell walls, was examined and modelled using Land Use Regression (LUR) models (de Rooij et al., 2018). While endotoxin serves as a useful marker of general bacterial emissions, further research is required to explore the spatial variation of other microbial agents more specific to livestock farming. In addition, due to the limited studies on microbial emission modelling, we seek to investigate whether the Random Forest (RF) method, which is inherently more flexible, could outperform LUR models. While machine learning models are increasingly being applied in the realm of chemical air pollution (Liu et al., 2022), there is a lack of studies focussing on their application for microbial pollution.

In this paper, our primary objective is to evaluate the feasibility of developing empirical spatial models for predicting residential exposure to livestock-specific microbial agents. Additionally, we aim to compare the performance of two modelling approaches: LUR and RF. We hypothesise that RF modelling, with its ability to handle nonlinearities and interactions between microbial emissions and predictor variables, will outperform standard linear regression-based modelling like LUR, resulting in more accurate estimates of microbial exposure. By employing these modelling approaches, we gain valuable insights into the patterns of livestock-related microbial emissions and its potential applicability in future health studies.

2. Materials and methods

2.1. Microbial air pollution data

This study made use of microbial air pollution data collected in the period of May 2014 to December 2015 as published by De Rooij et al (de Rooij et al., 2018; de Rooij et al., 2019a). In brief, airborne particle samples (PM_{10}) were collected using a filter-based technique (Teflon filters, Pall Corporation, Ann Arbor, USA) from 61 residential sites in the southeast of the Netherlands as part of the "Livestock Farming and Neighbouring Residents' Health" (VGO) project (de Rooij et al., 2018). Sites were selected to represent a variety of livestock-related characteristics (Fig. 1 gives a geographical overview of the selected sites and surrounding farms). Each site was repeatedly measured three or four times over a period of 14 days across all seasons. Simultaneous measurements were performed at 10 sites per sampling session, giving a total of 235 air samples. Additionally, to account for the temporal variability of the airborne microbial concentrations, a reference site was installed and continuously sampled throughout. Harvard Impactors (Air Diagnostics and Engineering Inc., Naples, ME, USA) sampled air at a flow rate of 10 L/min for 15 min of each hour during the 14-day sampling period. After sampling, air filters were processed further in the lab for DNA extraction which was performed using the NucliSENS Magnetic bead DNA extraction kit (Biomerieux-diagnostics, Marcy l'Etoile, France) (see de Rooij et al., 2019a for full extraction details). Subsequently quantitative polymerase chain reaction (qPCR) analyses were used to quantify DNA sequences from *E. coli*, *Staphylococcus* spp., *tetW* and *mecA* genes in ambient PM_{10} fraction (primers used can be found in Table S1). DNA sequences from *E. coli*, *Staphylococcus* spp., *tetW* and *mecA* were found (above the limit of detection) in the majority of

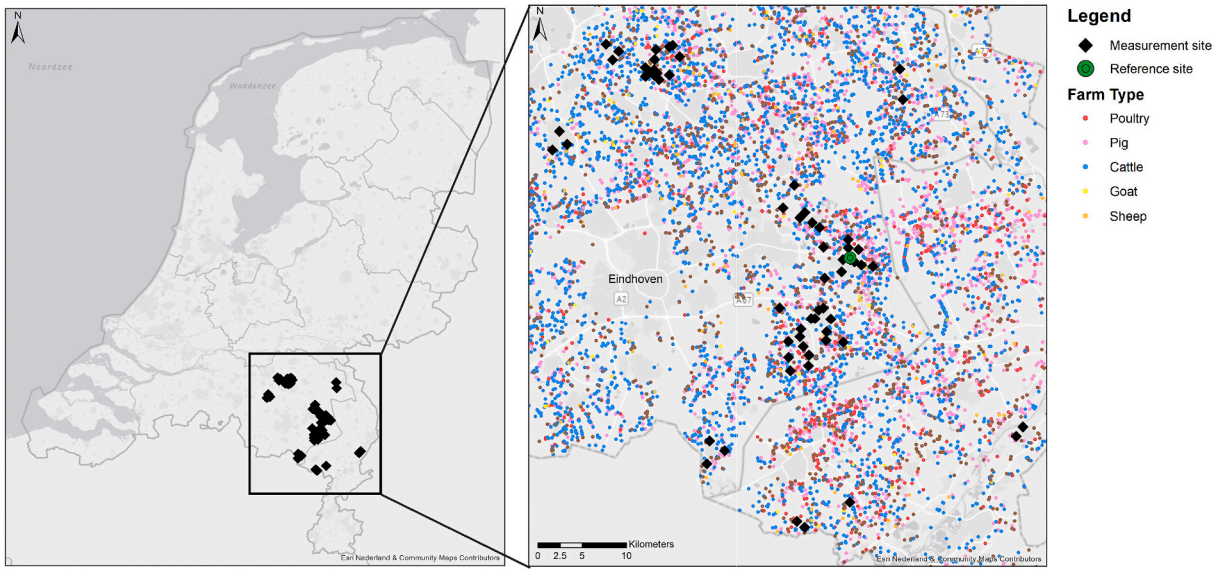


Fig. 1. Distribution of the 61 measurement sites in the Netherlands. The figure was created using ArcGIS (version 10.2.2; Esri) and the grey background is from the Esri Nederland & Community Maps Contributors.

samples (75%, 95%, 95% and 88%, respectively) (de Rooij et al., 2019a). All measured concentrations were adjusted for temporal variation by implementing the difference method using the concentrations measured at the reference site, as described previously (Eeftens et al., 2012; de Hoogh et al., 2013). Next, the annual average arithmetic mean concentration per site was calculated for each of the microbial agents, and these concentrations were subsequently natural logarithm (ln) transformed due to skewness of their distribution. Descriptive statistics of the annual average microbial concentrations can be found in Table S2, expressed both at the ln-transformed and original scales. To assess the efficacy of the difference method to adjust for temporal variation at each measurement site, we computed the per-site standard error of the mean (SEM). This allowed us to evaluate the accuracy of the annual average concentration at each site following correction for temporal variation. Formula S1 presents the SEM calculation utilised in this analysis.

2.2. Potential predictor variables

For this study, we made use of general and detailed livestock-related characteristics in the surroundings of the 61 measurement sites as computed with Geographic Information System (GIS) software (ArcGIS; version 10.2.2, Esri) using livestock data from 2015 as previously described (de Rooij et al., 2018; de Rooij et al., 2019a). These characteristics included the distance to the nearest farm and specific farm types, as well as the number of farms and animals within various buffer zones surrounding the measurement sites (250m, 500m, 1000m and 3000m). In addition, the distance-weighted number of farms and animals within these buffer zones were taken into account. For an overview of the GIS-computed livestock-related characteristics, see Table S3. The computed livestock-related predictors were winsorized to their 95th percentile, in accordance with de Rooij et al. (2018), in order to mitigate the influence of outliers due to their right-skewed distributions. To facilitate direct comparisons of the associations among different predictors, each variable was additionally scaled to its respective 10th to 90th percentile range. For model development, only livestock-related characteristics with non-zero values for more than 20 of the 61 sites were considered as predictors, resulting in a total of 133 potential predictors (refer to Table S4 for the list of potential predictors, including the descriptive statistics).

2.3. Model development

In this study we developed models to estimate the concentrations of *E. coli, Staphylococcus* spp., *tetW* and *mecA* using two different methods: 1) land use regression procedure (LUR) (Eeftens et al., 2012), and 2) random forest algorithm (RF) (Hu et al., 2017; Meng et al., 2018). Models were developed using the ln-transformed estimated yearly average concentrations of the microbial agents as dependent variables, while the livestock-related characteristics were used as the predictor variables. Both model types were constructed at three levels, each incorporating livestock-related characteristics of increasing detail as predictor variables. Level 1 utilised generic livestock-related characteristics such as distance to nearest farm, without specifying the type of animal. Level 2 models included characteristics specific to the main different animal species (poultry, pigs, horses, cows and sheep). The most comprehensive level, level 3, incorporated farm type-specific characteristics, such as subtypes of cattle farms (dairy, meat). For a detailed list of predictor variables considered in each model level, refer to Table S5. All modelling was performed using R Statistical Software (version 4.2.2; R Core Team, 2022).

2.3.1. Land use regression (LUR) models

Predictor variables were entered into the model using a forward supervised stepwise selection procedure as previously described, where predictors with the a priori defined direction of effect were offered in the model building process (Eeftens et al., 2012). The first variable included in the model was that with the highest adjusted explained variance (R^2). Additional variables were included in the model in a stepwise manner if they improved the adjusted R^2. Any model variables with a p value >0.1 or a variance inflation factor >3 were excluded from the model. Other model assumptions, including normality and homoscedasticity of residuals were evaluated, in addition to a check for influential observations (Cook's distance >1).

2.3.2. Random forest (RF) models

To develop these models, the R package *ranger* (version 0.14.1) was used to train and calibrate the RF models (Wright and Ziegler, 2017). Since the performance of a RF model is highly sensitive to the values of its hyperparameters, it is crucial to select the optimal set of values in order to improve the prediction accuracy and generalisability of the model, as well as to reduce the risk of overfitting (Cutler et al., 2012). The R package *tuneRanger* (version 0.5) was used to optimise the hyperparameters of the RF models (Probst et al., 2019). This was done through a grid search procedure, tuning the hyperparameters "mtry" (number of variables split at each node), "sample.fraction" (number of observations to sample for each decision tree) and "min.node.size" (the minimal size of the terminal nodes). The grid search performed an exhaustive search across a range of potential hyperparameter values to determine the best combination of values based on the highest R^2 value achieved by the model. The identified hyperparameters were subsequently used to train the model. The resulting optimised hyperparameter values for all models can be found in Table S6.

2.4. Model composition comparison

The composition of the LUR models can be readily determined by examining the predictor variables included in the models and their corresponding coefficients, which provide insight into the importance of each predictor. In the case of RF models, we could assess the individual contributions of each predictor variable to the model performance by calculating the reduction in node impurity caused by each predictor. In RF regression models, such as those developed here, node impurity refers to the variance following a split on that variable. The most important variables achieve larger decreases in impurity, or in other words, greater increases in purity, by explaining more variance in the data. The decrease in impurity is calculated for each variable in every tree in the RF by summing up the impurity reduction as a result of splitting by that variable. These values, computed for each tree, are then averaged across all trees to give the mean decrease in impurity. Subsequently, these importance scores are normalised, enabling us to make comparisons of the importance scores across different models.

2.5. Model evaluation

To evaluate model robustness, both LUR and RF models underwent 10-fold CV. This process involved dividing the measurement sites into training (90%) and test sets (10%), resulting in the development of 10 cross-validated models for each model (Briggs et al., 1997). These models were then applied to predict microbial concentrations at the corresponding sites that were held out from the model building process. Predictions from these 10 models were regressed against the actual microbial concentrations to compute the 10-fold CV R^2 and RMSE values. For the RF models, we employed a 10-fold nested CV approach to simultaneously tune the hyperparameters of the models and estimate model performance (Krstajic et al., 2014). Consequently, the best hyperparameter combination was selected for training on the full dataset. It has been shown that the nested CV approach provides unbiased estimates of model performance (Varma and Simon, 2006). The full dataset was split 10 times into outer loop training and test sets. Subsequently the outer loop training set was further divided into 10 folds for which we performed inner loop CV using the *tuneRanger* package to tune the hyperparameters of the model as described above. For each of the 10

outer loops, we trained a RF model using the optimal hyperparameters as identified from the inner loop 10-fold CV to make predictions at the sites held out from the model training. For all exposure models, training model R^2 and RMSE were computed by training the models on the full dataset and regressing the model predictions against the measured microbial concentrations for all 61 measurement sites. Refer to Formulae S2 and S3 for those implemented for the calculations of R^2 and RMSE values (10-fold CV and training), respectively.

3. Results

3.1. Microbial air pollution data

Considerable spatial variation of the annual average concentrations of the four microbial agents was observed (coefficients of variation for the absolute concentrations: 77.6%, 91.6%, 83.8%, 129.8% for *E. coli*, *Staphylococcus* spp., *tetW* and *mecA*, respectively). Fig. 2 shows the distributions of the temporally adjusted ln-transformed annual average concentrations of each microbial agent across the 61 measurement sites along with the coefficient of variation of the natural logarithm for each agent (CVnL). Moderate to strong correlations ($0.66 \leq \rho \leq 0.81$) were found between measured average concentrations of the four microbial agents (de Rooij et al., 2019a).

The efficacy of the difference method in adjusting for temporal variation at each measurement site is shown by the site-specific SEM both before and after adjustment. The mean site SEM values prior to temporal variation adjustment for *E. coli*, *Staphylococcus* spp., *tetW* and *mecA* were as follows: 0.50, 0.91, 0.66, 0.71 units, respectively, which all dropped to the following SEM values following temporal variation adjustment: 0.35, 0.61, 0.38, 0.54. These drops in SEM values showed us that by implementing the difference method, we improved the precision of the mean measurements by accounting for temporal variation. For *E. coli*, *Staphylococcus* spp. and *tetW*, the SEM values (post adjustment) were all less than 10% of the sample mean, and for *mecA* it was less than 20% of the sample mean, indicating a high level of accuracy in estimating the annual mean after correcting for temporal variation. Descriptive statistics of the SEM values before and after correction for temporal variation can be found in Table S7.

3.2. Model structure

The predictors selected in all three levels of the LUR modelling for each microbial agent along with their estimates are shown in Table S8. The LUR1 models for each microbial agent incorporated only one predictor variable, with the most common predictor being the number of livestock farms (irrespective of the animal type) weighted by the distance in a 3000m buffer (Σ(N/m)). The number of predictors included in each of the LUR2 models ranged between two and four, while for the LUR3 models, it ranged between three and five. LUR2 and LUR3 models for all microbial agents were dominated by predictors related to poultry, pigs and horses.

For all microbial agents, the most important predictor variable included in the RF1 models (based on reduction in node impurity) was the number of farms (all) in a 3000m buffer. The structure of the RF1 models for all microbial agents was comparable. For the RF2 models, predictors related to poultry and pigs dominated the top five predictors across the four microbial agents. When comparing the structure of the RF2 models for the different agents, we found that livestock predictors related to pig farms dominated most strongly in the *E. coli* RF2 models compared to the other agents, whereas livestock predictors related to poultry farms consistently appeared in the top five predictors for all microbial agents. Cattle farm predictors appeared in the top five only for *Staphylococcus* spp. and *mecA* RF2 models. In the RF3 models, predictors related to specific types of pigs and poultry were identified as variables of importance. As with the RF2 models, pig farm predictors were the most dominant in the *E. coli* RF3 model compared to the other agents, and did not appear in the top five predictors in the *Staphylococcus* spp. and *mecA* RF3 models. Poultry farm predictors consistently dominated

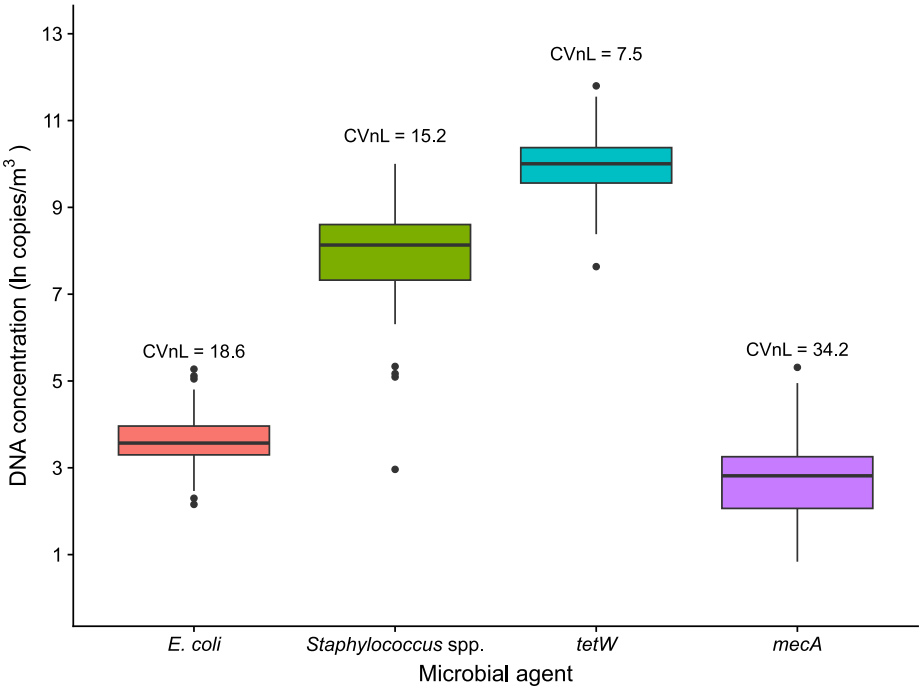


Fig. 2. Distribution of the temporally adjusted, ln-transformed annual average airborne concentrations of *E. coli*, *Staphylococcus* spp., *tetW* and *mecA*. The top and bottom limits of the boxes are the upper and lower quartiles of the mean pollutant concentration at each site. The horizontal line in each box is the median site average concentration and the whiskers indicate the variability outside the upper and lower quartiles. Individual circles plotted outside the whiskers are outliers defined as values which are either 1.5 × IQR above the third quartile or 1.5 × IQR below the first quartile.
CVnL = coefficient of variation of the natural logarithm for each agent.

in all RF3 models, and cattle farm predictors appeared in the top five predictors only for *Staphylococcus* spp. and *mecA* RF3 models. The list of the top five variables of importance in the three levels of RF models for each microbial agent can been found in Table S9.

Table 1 shows the predictor variables selected in the LUR2 models and the five most important variables in the RF2 models for each microbial agent. When comparing the structure of the LUR and RF models for each microbial agent, we observed notable similarity between the livestock predictors incorporated in the LUR models and the top five predictors in the corresponding RF model. In the *E. coli* models, predictor variables related to pigs were dominant. Conversely, models for the other livestock commensal, *Staphylococcus* spp., were dominated with predictor variables related to poultry, along with several predictors associated to horse farms. Models for *tetW* and *mecA* exhibited similar structures, with predictors related to poultry and pigs dominating these models, with a possible small additional contribution of horse farms. Most of the predictors incorporated in the LUR models, as well as those of high importance in the RF models, were livestock variables computed within a 3000m buffer, the largest buffer size considered in the model building process. Overall, for all microbial agents, livestock predictors related to poultry and pigs were incorporated the most frequently as predictors in the LUR models and with high importance in the RF models, indicating their likely importance in predicting spatial variation of these agents in this study.

3.3. Model performance

The model performance metrics are presented in Table 2 which includes the training and 10-fold CV R^2 and RMSE values for all LUR and RF models developed. Regarding the ability to predict spatial variation of the four microbial agents, the LUR models exhibited varying levels of explanatory power. Specifically, the best performing LUR model for each microbial agent could explain 37.0%, 10.0%, 8.4% and 16.7% of the spatial variation of *E. coli, Staphylococcus* spp., *tetW* and *mecA* at the hold-out sites, respectively. Overall, the LUR2 models, which incorporated predictors with greater livestock farm details than LUR1 but less detail than LUR3, performed the best in comparison to the LUR1 and LUR3 models. The RF models consistently outperformed their respective LUR model for all microbial agents. The best performing RF model for each microbial agent was able to explain 38.4%, 20.9%, 33.3% and 27.4% of the spatial variation of *E. coli, Staphylococcus* spp., *tetW* and *mecA* at the hold-out sites, respectively.

For all LUR and RF models, we observed a large drop from training model R^2 to 10-fold CV R^2. This drop was comparable for all LUR and RF models, with a mean decrease of 66.3% and 63.1% for LUR and RF models, respectively. This drop was more pronounced with increasingly complex models (i.e. models including more predictor variables).

3.4. Model predictions

We observed consistent agreements amongst the predictions from all LUR and RF training models for each microbial agent at the 61 monitoring sites, with strong and highly significant correlations (Pearson correlation coefficients (r) ranging from 0.62 to 0.99, all p values $<$ 0.001; Fig. 3). The average r amongst the predictions from the six models for *E. coli, Staphylococcus* spp., *tetW* and *mecA* were 0.879, 0.812, 0.804, 0.858, respectively. These results indicate that the different models for each microbial agent generated comparable predictions of their concentrations. Fig. 3 presents correlation matrices illustrating the relationships between the predictions from all LUR and RF models for each microbial agent individually. Besides these within-agent model agreements, also between-agent model prediction agreements were observed (Fig. S1). Modelled predictions of the different microbial pollutants exhibited moderate to good correlations, with r values ranging from 0.51 to 0.96, all p values $<$ 0.001. This underlines the similarities in model structure between the four microbial agents.

Table 1
List of predictors included in the LUR2 models and list of top five predictors included in the RF2 models for each microbial agent. LUR2 and RF2 models include animal species predictors but no subspecies information.

	E. coli	*Staphylococcus* spp.	tetW	mecA
LUR2	• N pigs weighted to distance in a 3000m buffer (Σ(N/m)) • N poultry in a 3000m buffer • N pigs in a 500m buffer • N horses in a 3000m buffer	• N poultry farms in a 3000m buffer • N horses in a 1000m buffer • N horse farms in a 3000m buffer	• N farms weighted to distance in a 3000m buffer (Σ(N/m)) • N poultry farms in a 1000m buffer • N horse farms in a 3000m buffer	• N poultry animals weighted to distance in a 3000m buffer (Σ(N/m)) • N farms (all) in 3000m buffer
RF2	• N pigs weighted to distance in a 3000m buffer (Σ(N/m)) • N pig farms weighted to distance in a 3000m buffer (Σ(N/m)) • N poultry in a 3000m buffer • N pig farms in a 3000m buffer • N farms (all) weighted to distance in a 3000m buffer (Σ(N/m))	• N farms (all) in a 3000m buffer • N poultry in a 3000m buffer • N cows in a 3000m buffer • N poultry farms weighted to distance in a 3000m buffer (Σ(N/m)) • Distance to the nearest pig farm (≥15 animals) (−1*m)	• N pig farms to distance in a 3000m buffer (Σ(N/m)) • N pigs weighted to distance in a 3000m buffer (Σ(N/m)) • N poultry weighted to distance in a 3000m buffer (Σ(N/m)) • N poultry in a 3000m buffer • N poultry weighted to distance in a 1000m buffer (Σ(N/m))	• N poultry in a 3000m buffer • N poultry weighted to distance in a 3000m buffer (Σ(N/m)) • N farms (all) in a 3000m buffer • N cows in a 3000m buffer • N poultry farms weighted to distance in a 3000m buffer (Σ(N/m))

Table 2
Training and 10-fold CV R^2 and RMSE for the three levels of LUR and RF models developed for each of the four microbial agents. Level 1 models are constructed using generic livestock-related characteristics without specifying the type of animal. Level 2 models include characteristics specific to the main different animal species (poultry, pigs, horses, cows, sheep). Level 3 models include subtypes of species.

	E. coli				*Staphylococcus* spp.				tetW				mecA			
	Training		10-fold CV		Training		10-fold CV		Training		10-fold CV		Training		10-fold CV	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
LUR1	0.353	0.543	0.163	0.631	0.223	1.042	0.078	1.157	0.242	0.643	0.080	0.725	0.255	0.810	0.120	0.898
LUR2	0.590	0.432	0.370	0.551	0.353	0.950	0.100	1.170	0.370	0.586	0.026	0.823	0.390	0.733	0.167	0.898
LUR3	0.654	0.397	0.199	0.642	0.406	0.911	0.090	1.181	0.439	0.553	0.084	0.776	0.471	0.682	0.143	0.933
RF1	0.530	0.483	0.276	0.576	0.865	0.523	0.209	1.056	0.383	0.593	0.158	0.679	0.424	0.728	0.185	0.847
RF2	0.865	0.311	0.384	0.536	0.714	0.756	0.168	1.081	0.937	0.257	0.333	0.607	0.686	0.595	0.264	0.807
RF3	0.820	0.357	0.365	0.547	0.664	0.798	0.145	1.097	0.813	0.397	0.313	0.618	0.768	0.535	0.274	0.802

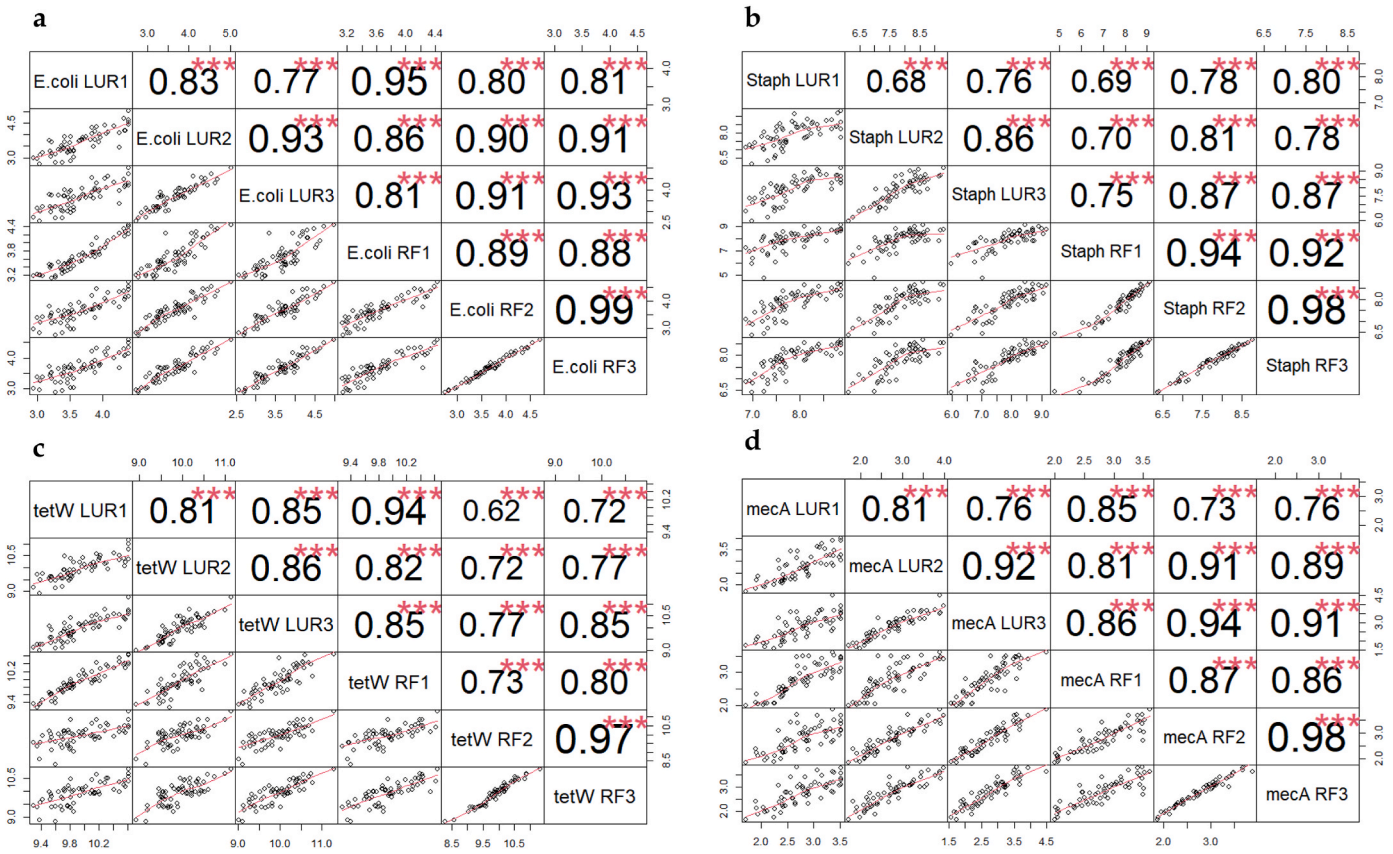


Fig. 3. Correlation matrix plot between all six training model predictions for **a** *E. coli*, **b** *Staphylococcus* spp., **c** *tetW* and **d** *mecA* concentrations. Numbers show Pearson correlation coefficients between the predictions from the different models, asterisks indicate the level of statistical significance for the corresponding correlation coefficients, where *** = p value < 0.001.

4. Discussion

The primary objective of our study was to determine whether empirical spatial models could be constructed and used to predict airborne concentrations of livestock-specific microbial agents in a livestock dense region. Furthermore, we aimed to assess and compare the performances of LUR and RF modelling to determine whether one method outperformed the other. Our findings demonstrated that these LUR and RF models, developed using comprehensive livestock information, are adequate at predicting spatial variation for each microbial agent, achieving moderate 10-fold CV performances. Additionally, we demonstrated that the RF models exhibited higher performances compared to LUR models, adding to the body of evidence comparing standard LUR methods to the machine learning-based RF approach (Brokamp et al., 2017; Brokamp et al., 2018; Chen et al., 2020).

As indicators of microbial emissions from livestock farms, we chose the microbial agents *E. coli, Staphylococcus* spp., *tetW* and *mecA* for this study. These agents are emitted by livestock farms and are suitable for characterisation in ambient air (de Rooij et al., 2019a). In addition, these agents exhibit a higher degree of specificity to livestock compared to the previously studied endotoxin, which could originate from a wider variety of sources emitting high quantities of (Gram-negative) bacteria (e.g. industrial waste recycling activities). In addition, clear spatial variation of these agents has been observed previously, highlighting their suitability for spatial modelling (de Rooij et al., 2019a). To our knowledge, only a limited number of previous studies have utilised empirical modelling techniques to predict bioaerosol concentrations in ambient air: de Rooij et al. (2018) for endotoxin in a rural environment and Hjort et al. (2016) for allergenic pollen in an urban environment. Using the same model build-up and 10-fold CV evaluation strategy, LUR models developed by de Rooij et al. (2018) could explain up to 32% of the spatial variation of endotoxin at the hold-out sites. This level of explanatory power is comparable to the findings observed in this study. These R^2 values are modest in contrast to modelling results of conventional air pollutants such as $PM_{2.5}$, NO_2 and O_3. Predicting bioaerosol concentrations in ambient air involves addressing a complex interplay of numerous contributing factors. Microorganisms, unlike chemical pollutants, engage in intricate interactions with each other, undergo growth and decay processes, and dynamically respond to environmental conditions. Furthermore, it is worth noting that the concentrations of bioaerosols are likely influenced not only by the number and type of farms (as taken into account in our modelling) but also by farm-level individual practices, including antimicrobial use, biosecurity measures, manure handling procedures, feeding practices, housing conditions and ventilation systems. Prior research has indicated that, beyond animal type and density, factors like antimicrobial use, biosecurity measures and feed and bedding type are linked to absolute and relative ARG abundances in pig and poultry farm dust (Luiken et al., 2022). Unfortunately, this detailed information was not available for our study.

Our LUR and RF models moderately explained the spatial variation in concentrations of the four microbial agents observed at the 61 measurement sites. The RF2 model for *E. coli* demonstrated the highest performance, with 38.4% of the spatial variation explained. Model robustness was evaluated using 10-fold CV, which has been shown by others to be a stringent validation test to gain insight into the generalisability of the models (Basagaña et al., 2012; Wang et al., 2012). Overall, the RF approach outperformed the LUR approach at all three model levels. This could be attributed to the capacity of RF models to capture non-linear relationships between predictors and outcomes, in addition to its enhanced ability to effectively capture interaction effects

among predictors, surpassing the capabilities of LUR models. Limited studies have made comparisons between the LUR and RF approaches in relation to air pollution and these previous studies were conducted in very different settings. Chen et al. (2019) developed spatial $PM_{2.5}$ models (based on 543 sites) and NO_2 models (based on 2399 sites) and found comparable performances between the LUR and RF algorithms in 5-fold CV. Kerckhoffs et al. (2019) found modest improvements in the spatial predictions of external measurements of ultrafine particles by RF models in comparison to LUR models.

Livestock predictors related to poultry and pigs appeared the most frequently overall across all LUR models and had the highest ranks of importance in the majority of RF models for all four microbial agents. It is important to note that these models are not designed specifically for source attribution and the selection of predictors in the models is not solely determined by their individual source strength, but also takes into account their geographical distribution. Despite similarities in the model structure for the microbial agents overall, we observed that *E. coli* models were dominated with predictor variables related to pig farms, *Staphylococcus* spp. models by those related to poultry farms and *tetW* and *mecA* models were dominated by both pig and poultry. These differences in dominance of predictor variables for the four microbial agents may reflect their source. *E. coli* is a bacterium commonly found in the gastrointestinal tracts of animals. Although it is usually a commensal bacterium, causing no disease in the animal, it can be transferred to the environment via faecal matter and possibly cause harm to others that ingest or inhale it (Ramos et al., 2020). It could be that pig faeces contain higher levels of *E. coli* compared to that of other livestock animals. Research has indicated that *E. coli* is one of the predominant genera in the gastrointestinal tract of adult pigs, whereas for chickens this is not the case (Forcina et al., 2022). Alternatively or additionally, it could be that specific pig farm-related practices, such as faecal storage methods, enhance the dispersion of *E. coli* in the air. *Staphylococcus* spp. have been identified in a variety of animals including poultry and pigs (Davis et al., 2018; Syed et al., 2020). Of particular note, bird feathers have been identified as a potential source of *Staphylococcus* spp. (Miskiewicz et al., 2018), hence providing a possible explanation to why poultry farms play an important role in explaining the spatial variation of *Staphylococcus* spp. concentrations. Many bacteria originating from both poultry and pig farms have been identified with ARGs. ARGs encoding resistance to tetracycline antibiotics (such as *tetW*) have been identified commonly in *E. coli* strains (Ramos et al., 2020). In addition, it has been shown that *Staphylococcus aureus* populations commonly harbour the *mecA* gene (Wielders et al., 2002). Despite this suggestion that pig and poultry farms may be the dominant players in contributing to microbial air pollution, the importance of pig and poultry farm variables in our models may also be caused by a higher contrast in geographical distribution across our study region of these farm types, allowing us to distinguish their association with microbial agents measured in ambient air more clearly. Livestock predictors with buffer sizes of 3000m were the most commonly incorporated in the LUR models and were of high importance in the RF models. This indicates that livestock-related microbial agents are likely to disperse several kilometres from farms. This implies that an individual's modelled microbial air pollutant concentrations at the home address is the cumulative exposure to those livestock farms in the surroundings that can best explain exposure contrast within this study population instead simply of the nearest farm contribution.

We observed large decreases from training R^2 to 10-fold CV R^2 values (mean decrease of 66.3% and 63.1% for LUR and RF models, respectively). Overall, as anticipated given the flexibility of the RF models, we observed that they exhibited higher training model R^2 values, which also corresponded to higher 10-fold CV R^2 values compared to the LUR models. A few previous studies have evaluated the effects of the number of measurement sites on LUR model performance (Basagaña et al., 2012; Wang et al., 2012; Johnson et al., 2010). These studies have demonstrated that training model R^2 values provide overly overoptimistic estimations of performance in models built with limited sample sizes, leading to increased disparities between the training and CV R^2 values. However, we observed even greater discrepancies than those reported in these studies. Johnson et al. (2010) and Basagaña et al. (2012) both demonstrated a convergence of training and validation R^2 values for LUR models developed for NO_2, benzene and $PM_{2.5}$ only when models were developed with 125 or more sites. Similar to our study, Wang et al. (2012) developed models using carefully selected measurements sites. They concluded that models developed with as few as 40 sites can provide reliable estimations, indicating the likely sufficiency of the number of monitoring sites included in our study. However, due to limited studies on LUR modelling for bioaerosols, it is difficult to assuredly determine the optimal number of measurement sites. Bioaerosols have been shown to have a higher natural variation than PM_{10} mass concentrations, and generally have a higher side-by-side variation (indicating both analytical and biological variability) compared to NO_2 concentrations, hence making extrapolation of these studies for bioaerosols difficult (de Rooij et al., 2018; de Rooij et al., 2019a). Challenges in collecting and detecting bioaerosols also augment the difficulty and limitations of including many sites within the monitoring campaign. The discrepancy between training and 10-fold CV R^2 values observed in our study may partially be explained by overfitting as we observe an increased discrepancy as model complexity increases (Babyak, 2004; Craig et al., 2007). However, this discrepancy is evident across all three levels of models, including the simpler models. This suggests that overfitting may not be the primary contributing factor to the observed discrepancy. This higher discrepancy observed in our models compared to those developed for NO_2 may reflect the likelihood that these agents have a larger diversity of sources in comparison to NO_2. NO_2 emissions are primarily associated with traffic sources in urban areas which exhibit more consistent emission patterns. This discrepancy between training and 10-fold CV R^2 values is an important issue to consider in our study, as the aim of these models is to provide reliable exposure predictions for unmonitored sites (for example the residential addresses of health study participants) that have not been used for model training. Nevertheless, despite the drop in R^2 values, considering the significant consistency observed between the different model predictions and the suboptimal alternatives for livestock exposure estimation (using proxies), we still deem the performance of the 10-fold CV model to be sufficient for estimating residential exposure. We believe that these model predictions will be of valuable use for future epidemiological studies investigating health effects of exposure to air pollution from livestock farms.

Although 10-fold CV can provide reliable estimates of a model's prediction performance for observations within the training dataset, it is not capable of evaluating a model's predictive ability at locations that were not included in the model development process. Validation by means of an external dataset would be valuable to determine our models' capabilities of estimating agent exposures at sites with differing spatial characteristics, as this would provide us with more robust conclusions regarding external performance and transferability. While an external validation dataset was not available for this study, monitoring sites were carefully selected to ensure that the full range of predictor variables was captured for the study region. Also, we did not have information on manure storage and land application which may have an impact on microbial pollutant concentrations in ambient air, as we know that manure handling is associated with increased occupational exposure to dust and endotoxin (Basinas et al., 2014). In research thus far, LUR modelling in rural settings is underemployed. Before our models can be applied in other countries, thorough validation is essential as applicability face several challenges. These challenges arise due to substantial variations in farming practices worldwide and the likely limited availability of equivalent detailed and accurately geocoded livestock data in other areas. While our models offer an attractive alternative to previously employed exposure proxies such as distance to the nearest farm, they do not yield perfect estimations, as indicated by their moderate R^2 values. Dispersion modelling, which is based on a

deterministic approach, has previously been applied successfully, after an intensive development phase, to model livestock farm-emitted endotoxin with the aim of estimating residential exposure (de Rooij et al., 2019b). This opens up possibilities to also explore this approach for other microbial agents.

While the specific health effects of exposure to airborne *E. coli*, *Staphylococcus* spp. and ARGs have not yet been explicitly investigated, it is widely recognised that opportunistic zoonotic bacteria originating from livestock farms can be transmitted through the air. These bacteria have been shown to have the potential to infect and subsequently lead to the development of disease. The Q fever epidemic in the Netherlands is a telling example; rural residents became ill after inhalation of airborne *Coxiella burnetii* bacteria emitted from goat farms (de Rooij et al., 2016). *Staphylococcus aureus* is known to have the ability to form biofilms, which have been detected in the airways of patients with chronic lung diseases (Cullen and McClean, 2015; Shimizu et al., 2015). In a study conducted by White et al. (2020), it was demonstrated that when *Staphylococcus aureus* is aerosolised while attached to farm dust, it triggers an inflammatory response in human granulocytes during *in vitro* experiments. This indicates that exposure to airborne bacteria could be associated with deteriorations in lung function (White et al., 2020). Several studies have proposed direct airborne transmission of ARGs from farm dust to humans as an important route of transmission to humans (McEachran et al., 2015; Li et al., 2018; Mbareche et al., 2019; Bos et al., 2016; Dohmen et al., 2017). Additional studies have investigated the impact of livestock exposure on the faecal resistome and ARG carriage (Zomer et al., 2016; Sun et al., 2020; Van Gompel et al., 2020). Although these studies have not specifically examined transmission routes, they consistently indicate that livestock exposure is a likely determinant for human ARG carriage in the gut in occupational settings and for neighbouring residents to farms. The observed associations between ARG carriage and livestock exposure might arise from direct ingestion, but could also be attributed to the inhalation of airborne bacteria carrying these ARGs, followed by the subsequent swallowing of these bacteria. It is plausible that airborne exposure exerts a more significant impact on the microbiome and resistome of the airways. This is a currently understudied topic but highly interesting, particularly considering the combined exposure to multiple pollutants originating from livestock farms. The chemical and microbial emissions from such livestock farms are likely to impact the airways through various pathophysiological mechanisms (Yang et al., 2020; Albright and Goldstein, 1996; Bessac et al., 2008).

5. Conclusion

The models developed in this study have the potential to estimate residential exposure to livestock farms within the context of epidemiological research. Employing these modelled exposures has the potential to minimise exposure misclassification, a common concern when relying on exposure proxies. Through the quantification of exposure to microbial emissions from livestock farms, our models offer a valuable pathway to gain better insights into the mechanisms underlying the observed health effects. Further understanding in this area could inform public health policies, particularly regarding monitoring and intervention strategies for regions with high livestock density and high population density, such as in the Netherlands.

Associated content

The Supplementary Material contains tables and figures related to model construction and evaluation.

CRediT authorship contribution statement

Beatrice Cornu Hewitt: Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Lidwien A.M. Smit:** Conceptualization, Funding acquisition, Resources, Writing – review & editing. **Warner van Kersen:** Formal analysis, Software, Writing – review & editing. **Inge M. Wouters:** Investigation, Writing – review & editing. **Dick J.J. Heederik:** Writing – review & editing. **Jules Kerckhoffs:** Writing – review & editing. **Gerard Hoek:** Writing – review & editing. **Myrna M.T. de Rooij:** Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Resources, Supervision, Writing – review & editing.

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

Data will be made available on request.

Acknowledgments

We would like to thank the participating residents who cooperated with us in the study. We thank Isabella van Schothorst, Heike Schmitt, Ingmar Janse, Arno Swart, Karlijn Moonen, Erik van Deurssen, Jack Spithoven, Erik van Nunen, Nena Burger, Siegfried de Wind, Gerdit Greve and Rozemarijn van der Plaats for their work with sample collection, sample processing and/or laboratory analyses. Additionally we acknowledge the provinces of Noord-Brabant and Limburg for their provision of livestock data used in this study.

The Livestock Farming and Neighbouring Residents' Health (VGO) study was funded by the Ministry of Health, Welfare and Sports and the Ministry of Economic Affairs of Netherlands. The current study was funded by a Dutch Research Council (NWO) Aspasia grant to Lidwien A. M. Smit (015.014.067).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2024.123590.

References

Albright, J.F., Goldstein, R.A., 1996. Airborne pollutants and the immune System. Otolaryngol. Neck Surg. 114 (2), 232–238. https://doi.org/10.1016/S0194-59989670173-0.

Babyak, M.A., 2004. What you see may not Be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med. 11.

Basagaña, X., Rivera, M., Aguilera, I., Agis, D., Bouso, L., Elosua, R., Foraster, M., de Nazelle, A., Nieuwenhuijsen, M., Vila, J., Künzli, N., 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmos. Environ. 54, 634–642. https://doi.org/10.1016/j.atmosenv.2012.01.064.

Basinas, I, Sigsgaard, T, Erlandsen, M, Andersen, NT, Takai, H, Heederik, D, Omland, Ø, Kromhout, H, Schlünssen, V, 2014 Jul. Exposure-affecting factors of dairy farmers' exposure to inhalable dust and endotoxin. Ann Occup Hyg 58 (6), 707–723. https://doi.org/10.1093/annhyg/meu024.

Bessac, B.F., Sivula, M., von Hehn, C.A., Escalera, J., Cohn, L., Jordt, S.-E., 2008. TRPA1 is a major oxidant sensor in murine airway sensory neurons. J. Clin. Invest. 118 (5), 1899–1910. https://doi.org/10.1172/JCI34192.

Borlée, F., Yzermans, C.J., van Dijk, C.E., Heederik, D., Smit, L.A.M., 2015. Increased respiratory symptoms in COPD patients living in the vicinity of livestock farms. Eur. Respir. J. 46 (6), 1605–1614. https://doi.org/10.1183/13993003.00265-2015.

Bos, M.E.H., Verstappen, K.M., van Cleef, B.A.G.L., Dohmen, W., Dorado-García, A., Graveland, H., Duim, B., Wagenaar, J.A., Kluytmans, J.A.J.W., Heederik, D.J.J., 2016. Transmission through air as a possible route of exposure for MRSA. J. Expo. Sci. Environ. Epidemiol. 26 (3), 263–269. https://doi.org/10.1038/jes.2014.85.

Briggs, D.J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K., Van Der Veen, A., 1997. Mapping urban air pollution using GIS: a regression-based approach. Int. J. Geogr. Inf. Sci. 11 (7), 699–718. https://doi.org/10.1080/136588197242158.

Brokamp, C., Jandarov, R., Rao, M.B., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos. Environ. 151, 1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066.

Brokamp, C., Jandarov, R., Hossain, M., Ryan, P., 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. Environ. Sci. Technol. 52 (7), 4173–4179. https://doi.org/10.1021/acs.est.7b05381.

Chai, L., Kröbel, R., Janzen, H.H., Beauchemin, K.A., McGinn, S.M., Bittman, S., Atia, A., Edeogu, I., MacDonald, D., Dong, R., 2014. A regional mass balance model based on total ammoniacal nitrogen for estimating ammonia emissions from beef cattle in alberta Canada. Atmos. Environ. 92, 292–302. https://doi.org/10.1016/j.atmosenv.2014.04.037.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934 https://doi.org/10.1016/j.envint.2019.104934.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Weinmayr, G., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Atkinson, R., Janssen, N.A.H., Martin, R.V., Samoli, E., Andersen, Z.J., Oftedal, B.M., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Brunekreef, B., Hoek, G., 2020. Development of europe-wide models for particle elemental composition using supervised linear regression and random forest. Environ. Sci. Technol. 54 (24), 15698–15709. https://doi.org/10.1021/acs.est.0c06595.

Craig, M.H., Sharp, B.L., Mabaso, M.L., Kleinschmidt, I., 2007. Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. Int. J. Health Geogr. 6 (1), 44. https://doi.org/10.1186/1476-072X-6-44.

Cullen, L., McClean, S., 2015. Bacterial adaptation during chronic respiratory infections. Pathogens 4 (1), 66–89. https://doi.org/10.3390/pathogens4010066.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests. In: Zhang, C., Ma, Y. (Eds.), Ensemble Machine Learning. Springer New York, New York, NY, pp. 157–175. https://doi.org/10.1007/978-1-4419-9326-7_5.

Davis, M.F., Pisanic, N., Rhodes, S.M., Brown, A., Keller, H., Nadimpalli, M., Christ, A., Ludwig, S., Ordak, C., Spicer, K., Love, D.C., Larsen, J., Wright, A., Blacklin, S., Flowers, B., Stewart, J., Sexton, K.G., Rule, A.M., Heaney, C.D., 2018. Occurrence of Staphylococcus aureus in swine and swine workplace environments on industrial and antibiotic-free hog operations in North Carolina, USA: a one health pilot study. Environ. Res. 163, 88–96. https://doi.org/10.1016/j.envres.2017.12.010.

Dohmen, W., Schmitt, H., Bonten, M., Heederik, D., 2017. Air exposure as a possible route for ESBL in pig Farmers. Environ. Res. 155, 359–364. https://doi.org/10.1016/j.envres.2017.03.002.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dėdelė, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Gražulevičienė, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Nádor, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M.-Y., Yli-Tuomi, T., Varró, M.J., Vienneau, D., Klot, S. von, Wolf, K., Brunekreef, B., Hoek, G., 2012. Development of land use regression models for PM $_{2.5}$, PM $_{2.5}$ absorbance, PM $_{10}$ and PM $_{coarse}$ in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46 (20), 11195–11205. https://doi.org/10.1021/es301948k.

Forcina, G., Pérez-Pardal, L., Carvalheira, J., Beja-Pereira, A., 2022. Gut microbiome studies in livestock: achievements, challenges, and perspectives. Animals 12 (23), 3375. https://doi.org/10.3390/ani12233375.

Franceschini, G., Bottino, M., Millet, I., Martello, E., Zaltron, F., Favretto, A., Vonesch, N., Tomao, P., Mannelli, A., 2019. Assessment of the exposure of Turkey Farmers to antimicrobial resistance associated with working practices. Vet. Sci. 6 (1), 13. https://doi.org/10.3390/vetsci6010013.

Gao, F.-Z., He, L.-Y., Bai, H., He, L.-X., Zhang, M., Chen, Z.-Y., Liu, Y.-S., Ying, G.-G., 2023. Airborne bacterial community and antibiotic resistome in the swine farming environment: metagenomic insights into livestock relevance, pathogen hosts and public risks. Environ. Int. 172, 107751 https://doi.org/10.1016/j.envint.2023.107751.

Gibbs, S.G., Green, C.F., Tarwater, P.M., Mota, L.C., Mena, K.D., Scarpino, P.V., 2006. Isolation of antibiotic-resistant bacteria from the air plume downwindof a swine confined or concentrated animal feeding operation. Environ. Health Perspect. 114 (7), 1032–1037. https://doi.org/10.1289/ehp.8910.

Van Gompel, L., Luiken, R.E.C., Hansen, R.B., Munk, P., Bouwknegt, M., Heres, L., Greve, G.D., Scherpenisse, P., Jongerius-Gortemaker, B.G.M., Tersteeg-Zijderveld, M. H.G., García-Cobos, S., Dohmen, W., Dorado-García, A., Wagenaar, J.A., Urlings, B. A.P., Aarestrup, F.M., Mevius, D.J., Heederik, D.J.J., Schmitt, H., Bossers, A., Smit, L. A.M., 2020. Description and determinants of the faecal resistome and microbiome of Farmers and slaughterhouse workers: a metagenome-wide cross-sectional study. Environ. Int. 143, 105939 https://doi.org/10.1016/j.envint.2020.105939.

Hjort, J., Hugg, T.T., Antikainen, H., Rusanen, J., Sofiev, M., Kukkonen, J., Jaakkola, M. S., Jaakkola, J.J.K., 2016. Fine-scale exposure to allergenic pollen in the urban environment: evaluation of land use regression approach. Environ. Health Perspect. 124 (5), 619–626. https://doi.org/10.1289/ehp.1509761.

de Hoogh, K., Wang, M., Adam, M., Badaloni, C., Beelen, R., Birk, M., Cesaroni, G., Cirach, M., Declercq, C., Dėdelė, A., Dons, E., de Nazelle, A., Eeftens, M., Eriksen, K., Eriksson, C., Fischer, P., Gražulevičienė, R., Gryparis, A., Hoffmann, B., Jerrett, M., Katsouyanni, K., Iakovides, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Mosler, G., Nádor, G., Nieuwenhuijsen, M., Pershagen, G., Peters, A., Phuleria, H., Probst-Hensch, N., Raaschou-Nielsen, O., Quass, U., Ranzi, A., Stephanou, E., Sugiri, D., Schwarze, P., Tsai, M.-Y., Yli-Tuomi, T., Varró, M.J., Vienneau, D., Weinmayr, G., Brunekreef, B., Hoek, G., 2013. Development of land use regression models for particle composition in twenty study areas in europe. Environ. Sci. Technol. 47 (11), 5778–5786. https://doi.org/10.1021/es400156t.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM $_{2.5}$ concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51 (12), 6936–6944. https://doi.org/10.1021/acs.est.7b01210.

Jaarverslag, 2015. Autoriteit Diergeneesmiddelen (SDa), 2016. https://www.autoriteitdiergeneesmiddelen.nl/nl/publicaties/jaarverslagen.

Johnson, M., Isakov, V., Touma, J.S., Mukerjee, S., Özkaynak, H., 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmos. Environ. 44 (30), 3660–3668. https://doi.org/10.1016/j.atmosenv.2010.06.041.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ. Sci. Technol. 53 (3), 1413–1421. https://doi.org/10.1021/acs.est.8b06038.

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminf. 6 (1), 10. https://doi.org/10.1186/1758-2946-6-10.

Li, J., Cao, J., Zhu, Y., Chen, Q., Shen, F., Wu, Y., Xu, S., Fan, H., Da, G., Huang, R., Wang, J., de Jesus, L, A., Morawska, L., Chan, C.K., Peccia, J., Yao, M., 2018. Global survey of antibiotic resistance genes in air. Environ. Sci. Technol. 52 (19), 10975–10984. https://doi.org/10.1021/acs.est.8b02204.

Liu, X., Lu, D., Zhang, A., Liu, Q., Jiang, G., 2022. Data-driven machine learning in environmental pollution: gains and problems. Environ. Sci. Technol. 56 (4), 2124–2133. https://doi.org/10.1021/acs.est.1c06157.

Luiken, R.E.C., Heederik, D.J.J., Scherpenisse, P., Van Gompel, L., van Heijnsbergen, E., Greve, G.D., Jongerius-Gortemaker, B.G.M., Tersteeg-Zijderveld, M.H.G., Fischer, J., Juraschek, K., Skarżyńska, M., Zając, M., Wasyl, D., Wagenaar, J.A., Smit, L.A.M., Wouters, I.M., Mevius, D.J., Schmitt, H., 2022. Determinants for antimicrobial resistance genes in farm dust on 333 poultry and pig farms in nine European countries. Environ. Res. 208, 112715 https://doi.org/10.1016/j.envres.2022.112715.

McEachran, A.D., Blackwell, B.R., Hanson, J.D., Wooten, K.J., Mayer, G.D., Cox, S.B., Smith, P.N., 2015. Antibiotics, bacteria, and antibiotic resistance genes: aerial transport from cattle feed yards via particulate matter. Environ. Health Perspect. 123 (4), 337–343. https://doi.org/10.1289/ehp.1408555.

Meng, X., Hand, J.L., Schichtel, B.A., Liu, Y., 2018. Space-time trends of PM2.5 constituents in the conterminous United States estimated by a machine learning approach, 2005–2015. Environ. Int. 121, 1137–1147. https://doi.org/10.1016/j.envint.2018.10.029.

Mevius, D., Heederik, D., 2014. Reduction of antibiotic use in animals “let’s go Dutch.”. J. Für Verbraucherschutz Leb. 9 (2), 177–181. https://doi.org/10.1007/s00003-014-0874-z.

Mirabelli, M.C., Wing, S., Marshall, S.W., Wilcosky, T.C., 2006. Asthma symptoms among adolescents who attend public schools that are located near confined swine feeding operations. Pediatrics 118 (1), e66–e75. https://doi.org/10.1542/peds.2005-2812.

Miskiewicz, A., Kowalczyk, P., Oraibi, S.M., Cybulska, K., Misiewicz, A., 2018. Bird feathers as potential sources of pathogenic microorganisms: a new look at old diseases. Antonie Leeuwenhoek 111 (9), 1493–1507. https://doi.org/10.1007/s10482-018-1048-2.

Pavilonis, B.T., Sanderson, W.T., Merchant, J.A., 2013. Relative exposure to swine animal feeding operations and childhood asthma prevalence in an agricultural cohort. Environ. Res. 122, 74–80. https://doi.org/10.1016/j.envres.2012.12.008.

Pohl, H.R., Citra, M., Abadin, H.A., Szadkowska-Stańczyk, I., Kozajda, A., Ingerman, L., Nguyen, A., Murray, H.E., 2017. Modeling emissions from CAFO poultry farms in Poland and evaluating potential risk to surrounding populations. Regul. Toxicol. Pharmacol. 84, 18–25. https://doi.org/10.1016/j.yrtph.2016.11.005.

Probst, P., Wright, M.N., Boulesteix, A., 2019. Hyperparameters and tuning strategies for random forest. WIREs Data Min. Knowl. Discov. 9 (3) https://doi.org/10.1002/widm.1301.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL Https://Www.R-Project.Org/.

Radon, K., Schulze, A., Ehrenstein, V., van Strien, R.T., Praml, G., Nowak, D., 2007. Environmental exposure to confined animal feeding operations and respiratory health of neighboring residents. Epidemiology 18 (3), 300–308. https://doi.org/10.1097/01.ede.0000259966.62137.84.

Ramos, S., Silva, V., Dapkevicius, M. de L.E., Caniça, M., Tejedor-Junco, M.T., Igrejas, G., Poeta, P., 2020. Escherichia coli as commensal and pathogenic bacteria among food-producing animals: health implications of extended spectrum β-lactamase (ESBL) production. Animals 10 (12), 2239. https://doi.org/10.3390/ani10122239.

Robles-Jimenez, L.E., Aranda-Aguirre, E., Castelan-Ortega, O.A., Shettino-Bermudez, B. S., Ortiz-Salinas, R., Miranda, M., Li, X., Angeles-Hernandez, J.C., Vargas-Bello-Pérez, E., Gonzalez-Ronquillo, M., 2021. Worldwide traceability of antibiotic residues from livestock in wastewater and soil: a systematic review. Animals 12 (1), 60. https://doi.org/10.3390/ani12010060.

de Rooij, M.M.T., Borlée, F., Smit, L.A.M., de Bruin, A., Janse, I., Heederik, D.J.J., Wouters, I.M., 2016. Detection of Coxiella burnetii in ambient air after a large Q fever outbreak. PLoS One 11 (3), e0151281. https://doi.org/10.1371/journal.pone.0151281.

de Rooij, M.M.T., Heederik, D.J.J., van Nunen, E.J.H.M., van Schothorst, I.J., Maassen, C. B.M., Hoek, G., Wouters, I.M., 2018. Spatial variation of endotoxin concentrations measured in ambient PM10 in a livestock-dense area: implementation of a land-use regression approach. Environ. Health Perspect. 126 (1), 017003 https://doi.org/10.1289/EHP2252.

de Rooij, M.M.T., Hoek, G., Schmitt, H., Janse, I., Swart, A., Maassen, C.B.M., Schalk, M., Heederik, D.J.J., Wouters, I.M., 2019a. Insights into livestock-related microbial concentrations in air at residential level in a livestock dense area. Environ. Sci. Technol. 53 (13), 7746–7758. https://doi.org/10.1021/acs.est.8b07029.

de Rooij, M.M.T., Smit, L.A.M., Erbrink, H.J., Hagenaars, T.J., Hoek, G., Ogink, N.W.M., Winkel, A., Heederik, D.J.J., Wouters, I.M., 2019b. Endotoxin and particulate matter emitted by livestock farms and respiratory health effects in neighboring residents. Environ. Int. 132, 105009 https://doi.org/10.1016/j.envint.2019.105009.

Schinasi, L., Horton, R.A., Guidry, V.T., Wing, S., Marshall, S.W., Morland, K.B., 2011. Air pollution, lung function, and physical symptoms in communities near concentrated swine feeding operations. Epidemiology 22 (2), 208–215. https://doi.org/10.1097/EDE.0b013e3182093c8b.

Schulze, A., Römmelt, H., Ehrenstein, V., van Strien, R., Praml, G., Küchenhoff, H., Nowak, D., Radon, K., 2011. Effects on pulmonary health of neighboring residents of concentrated animal feeding operations: exposure assessed using optimized estimation technique. Arch. Environ. Occup. Health 66 (3), 146–154. https://doi.org/10.1080/19338244.2010.539635.

Shimizu, K., Yoshii, Y., Morozumi, M., Chiba, N., Ubukata, K., Uruga, H., Hanada, S., Saito, N., Kadota, T., Wakui, H., Ito, S., Takasaka, N., Minagawa, S., Kojima, J., Numata, T., Hara, H., Kawaishi, M., Saito, K., Araya, J., Kaneko, Y., Nakayama, K., Kishi, K., Kuwano, K., 2015. Pathogens in COPD exacerbations identified by comprehensive real-time PCR plus older methods. Int. J. Chronic Obstr. Pulm. Dis. 2009 https://doi.org/10.2147/COPD.S82752.

Sigurdarson, S.T., Kline, J.N., 2006. School proximity to concentrated animal feeding operations and prevalence of asthma in students. Chest 129 (6), 1486–1491. https://doi.org/10.1378/chest.129.6.1486.

Sun, J., Liao, X.-P., D'Souza, A.W., Boolchandani, M., Li, S.-H., Cheng, K., Luis Martínez, J., Li, L., Feng, Y.-J., Fang, L.-X., Huang, T., Xia, J., Yu, Y., Zhou, Y.-F., Sun, Y.-X., Deng, X.-B., Zeng, Z.-L., Jiang, H.-X., Fang, B.-H., Tang, Y.-Z., Lian, X.-L., Zhang, R.-M., Fang, Z.-W., Yan, Q.-L., Dantas, G., Liu, Y.-H., 2020. Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. Nat. Commun. 11 (1), 1427. https://doi.org/10.1038/s41467-020-15222-y.

Syed, M.A., Ullah, H., Tabassum, S., Fatima, B., Woodley, T.A., Ramadan, H., Jackson, C. R., 2020. Staphylococci in poultry intestines: a comparison between farmed and household chickens. Poultry Sci. 99 (9), 4549–4557. https://doi.org/10.1016/j.psj.2020.05.051.

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinf. 7 (1), 91. https://doi.org/10.1186/1471-2105-7-91.

de Vries, W., Kros, J., Voogd, J.C., Ros, G.H., 2023. Integrated assessment of agricultural practices on large scale losses of ammonia, greenhouse gases, nutrients and heavy metals to air and water. Sci. Total Environ. 857, 159220 https://doi.org/10.1016/j.scitotenv.2022.159220.

Wang, M., Beelen, R., Eeftens, M., Meliefste, K., Hoek, G., Brunekreef, B., 2012. Systematic evaluation of land use regression models for NO $_2$. Environ. Sci. Technol. 46 (8), 4481–4489. https://doi.org/10.1021/es204183v.

White, J.K., Nielsen, J.L., Larsen, C.M., Madsen, A.M., 2020. Impact of dust on airborne Staphylococcus aureus' viability, culturability, inflammogenicity, and biofilm forming capacity. Int. J. Hyg Environ. Health 230, 113608. https://doi.org/10.1016/j.ijheh.2020.113608.

Wielders, C.L.C., Fluit, A.C., Brisse, S., Verhoef, J., Schmitz, F.J., 2002. *MecA* gene is widely disseminated in *Staphylococcus aureus* population. J. Clin. Microbiol. 40 (11), 3970–3975. https://doi.org/10.1128/JCM.40.11.3970-3975.2002.

Wright, M.N., Ziegler, A. Ranger, 2017. A fast implementation of random forests for high dimensional data in *C++* and *R*. J. Stat. Software 77 (1). https://doi.org/10.18637/jss.v077.i01.

Yang, J., Kim, E.K., Park, H.J., McDowell, A., Kim, Y.-K., 2020. The impact of bacteria-derived ultrafine dust particles on pulmonary diseases. Exp. Mol. Med. 52 (3), 338–347. https://doi.org/10.1038/s12276-019-0367-3.

Zomer, T.P., Wielders, C.C.H., Veenman, C., Hengeveld, P., van der Hoek, W., de Greeff, S.C., Smit, L.A.M., Heederik, D.J., Yzermans, C.J., Bosch, T., Maassen, C.B.M., van Duijkeren, E., 2016. MRSA in persons not living or working on a farm in a livestock-dense area: prevalence and risk factors. J. Antimicrob. Chemother. dkw483. https://doi.org/10.1093/jac/dkw483.

Mbareche H, Veillette M, Pilote J, Létourneau V, Duchaine C. Bioaerosols Play a Major Role in the Nasopharyngeal Microbiota Content in Agricultural Environment. *Int J Environ Res Public Health*. 2019;16(8):1375. Published 2019 Apr 16. doi:10.3390/ijerph16081375.