pagepress

Geospatial Health 2017; volume 12:539

Assessing effects of structural zeros on models of canine cancer incidence: a case study of the Swiss Canine Cancer Registry

Gianluca Boo,[1,2] Stefan Leyk,[3] Sara Irina Fabrikant,[1] Andreas Pospischil,[2] Ramona Graf[2]

[1]Department of Geography, University of Zurich, Zurich; [2]Collegium Helveticum, University of Zurich and Swiss Federal Institute of Technology in Zurich, Zurich, Switzerland; [3]Department of Geography, University of Colorado, Boulder, CO, USA

Abstract

Epidemiological research of canine cancers could inform comparative studies of environmental determinants for a number of human cancers. However, such an approach is currently limited because canine cancer data sources are still few in number and often incomplete. Incompleteness is typically due to under-ascertainment of canine cancers. A main reason for this is because dog owners commonly do not seek veterinary care for this diagnosis. Deeper knowledge on under-ascertainment is critical for modelling canine cancer incidence, as an indication of zero incidence might originate from the sole absence of diagnostic examinations within a given sample unit. In the present case study, we investigated effects of such structural zeros on models of canine cancer incidence. In doing so, we contrasted two scenarios for modelling incidence data retrieved from the Swiss Canine Cancer Registry. The first scenario was based on the complete enumeration of incidence data for all Swiss municipal units. The second scenario was based on a filtered sample that systematically discarded structural zeros in those municipal units where no diagnostic examination had been performed. By means of cross-validation, we assessed and contrasted statistical performance and predictive power of the two modelling scenarios. This analytical step allowed us to demonstrate that structural zeros impact on the generalisability of the model of canine cancer incidence, thus challenging future comparative studies of canine and human cancers. The results of this case study show that increased awareness about the effects of structural zeros is critical to epidemiological research.

Correspondence: Gianluca Boo, Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.
Tel: +41.44.635.51.54.
E-mail: gianluca.boo@geo.uzh.ch

Acknowledgements: this study is funded by the Collegium Helveticum, a joint initiative of the University of Zurich and the Swiss Federal Institute of Technology in Zurich, through the grant accorded to its fellows Andreas Pospischil and Kay W. Axhausen.

Key words: Canine cancer registries; Under-ascertainment; Structural zeros; Regression analysis; Cross validation.

Received for publication: 22 December 2016.
Accepted for publication: 18 April 2017.

Geospatial Health 2017; 12:539
doi:10.4081/gh.2017.539

Introduction

Epidemiological research can provide insights into demographic and environmental determinants of canine cancers, a group of degenerative diseases listed among the leading causes of death in dogs (Vail and MacEwen, 2000; Pinho *et al*., 2012). For instance, studies have indicated that breed, sex, and age are important demographic determinants of several canine cancers (Bronson, 1982; Eichelberg and Seine, 1996; Lund *et al*., 1999; Michell, 1999; Proschowsky *et al*., 2003). Studies have also shown that a number of canine cancers can be linked to specific environmental determinants, such as exposure to tobacco smoke (Reif *et al*., 1998), combustion products (Bukowski *et al*., 1998), herbicides (Hayes *et al*., 1981), insecticides (Glickman *et al*., 1989), asbestos (Glickman *et al*., 1983), as well as paints and solvents (Gavazza *et al*., 2001). As such exposures mostly occur within a living environment shared with the owner, epidemiological research of canine cancers might also inform comparative studies of environmental determinants for human cancers, for instance, in the bladder (Hayes *et al*., 1981; Glickman *et al*., 1989), respiratory tract (Bukowski *et al*., 1998; Reif *et al*., 1998) and mammary gland (Owen, 1979; Vascellari *et al*., 2016). Despite the recognised advantages of such comparative studies (Schmidt, 2009; Scotch *et al*., 2009; Reif, 2011), this approach is currently limited because canine cancer data sources are still few in number and often incomplete (Brønden *et al*., 2007; Nødtvedt *et al*., 2012). The incompleteness of existing data sources is often a result of under-reporting and under-ascertainment of canine cancers (Brønden *et al*., 2007; Nødtvedt *et al*., 2012). Under-reporting occurs when the result of a performed diagnostic examination is not reported in any data source (Gibbons *et al*., 2014), while under-ascertainment occurs when the diagnostic examination has not been performed at all, because the dog's owner did not seek veterinary care for a diagnosis (Gibbons *et al*., 2014). Under-ascertainment thus implies that the information about the missing diagnostic examination cannot be retrieved from the data source, and this issue cannot be addressed through the imputation meth-

ods employed for assessing under-reporting (Gibbons *et al*., 2014; Dvorzak and Wagner, 2016). As a consequence, under-ascertainment is often estimated as a random effect within statistical frameworks for modelling incidence data (Gibbons *et al*., 2014; Dvorzak and Wagner, 2016).

When modelling canine cancer incidence, under-ascertainment might become critical, because an indication of zero incidences can originate from the sole absence of diagnostic examinations within a given sample unit (Hu *et al*., 2011; He *et al*., 2014). Such zeros are the manifestation of a structural phenomenon in the data source and should therefore be discarded from any effort at modelling the incidence data (Hu *et al*., 2011; He *et al*., 2014). Still, this sort of structural zeros are difficult to discard because they are often mistaken for sampling zeros, which result from diagnostic examinations performed within a sample unit (Mohri and Roark, 2005; Legendre and Legendre, 2012). Due to the persistent uncertainty surrounding the nature of zero incidences, little is known about the effects of structural zeros on models of canine cancer incidence. To fill this important knowledge gap, in the present case study investigated the effects of structural zeros on models of canine cancer incidence retrieved across Swiss municipalities in 2008, thus addressing the potential under-ascertainment affecting the Swiss Canine Cancer Registry (Grüntzig *et al*., 2015, 2016). In doing so, we contrast two modelling scenarios. The first scenario was based on the complete enumeration of incidence data for all municipal units. The second scenario was based on a filtered sample that systematically discarded structural zeros: the municipal units where no diagnostic examination has been performed. This filtering step was made possible by the exceptionally rich attribution of the Swiss Canine Cancer Registry, which contains information on the number of diagnostic examinations performed within each municipal unit in a specific year (Grüntzig *et al*., 2015, 2016). By contrasting the statistical performance and predictive power of these two modelling scenarios in a cross-validation framework (Snee, 1977; Picard and Cook, 1984), we provided new insights into the effects of structural zeros in models of canine cancer incidence, and highlight challenges for future comparative studies of canine and human cancers (Kukull and Ganguli, 2012; St. Sauver *et al*., 2012).

Materials and Methods

Canine cancer diagnostic examinations and demographic risk factors

The Swiss Canine Cancer Registry is a unique data source for epidemiological research of canine cancers, as it comprises more than 120,000 diagnostic examinations performed in Switzerland between 1955 and 2008 through necropsy, biopsy, and cytology tests (Grüntzig *et al*., 2015, 2016). This data source has been retrospectively assembled by the Collegium Helveticum Zurich for future comparative studies of canine and human cancers, and it is currently in the process of being updated to include diagnostic examinations for the most recent years (Grüntzig *et al*., 2015, 2016). As previous research suggests that the accuracy and completeness of the Swiss Canine Cancer Registry sensibly decrease in earlier years (Grüntzig *et al*., 2015), we only retrieved the 7057 diagnostic examinations performed in 2008. These examinations allowed for the ascertainment of 3509 canine cancer cases. For the purpose of the present case study, we enumerated the number of diagnostic examinations and observed canine cancer cases at the municipal level. In addition, to account for determinants involving the demographic characteristics of the at-risk canine population, we accessed demographic information on the 496,689 dogs living in Switzerland in 2008. This information was retrieved from the Swiss canine population database, compiled by Animal Identity Service AG following the legal obligation of dog microchipping and registration established in Switzerland in 2006 (Pospischil *et al*., 2013; ANIS, 2016). For previous years, demographic information can be retrieved only for a limited number of municipalities, generally located in urban areas, or as estimates at the country level (Pospischil *et al*., 2013). We derived the size of the at-risk population, and the average age and sex ratio of dogs within municipalities because these variables are important demographic determinants for the biological predisposition to several canine cancers (Bronson, 1982; Eichelberg and Seine, 1996; Lund *et al*., 1999; Michell, 1999; Proschowsky *et al*., 2003).

Urban character, socioeconomic status and distance to veterinary care

We assessed the urban character and socio-economic status of Swiss municipalities, as existing studies suggest that these are the key variables for characterising potential under-ascertainment of canine cancers (Boo *et al*., 2015, 2016). We estimated the urban character as human population densities at the municipal level, using the extent of residential land within municipalities as the areal denominator. For this purpose, we used the Swiss Federal Statistical Office census data for 2008 (SFSO, 2016) and information on the areal extent of residential land derived from the building and dwelling survey conducted by the Swiss Federal Statistical Office in 2014 (SFSO, 2016). The socio-economic status is approximated based on average national income tax information collected by the Swiss Federal Tax Administration in 2008 (SFTA, 2016). We also derived the travel distance to veterinary care within municipalities from a hectometric raster representing distance along roads (Delamater *et al*., 2012). In doing so, we assumed that increasing travel distance to veterinary services would be an important determinant for potential under-ascertainment (Boo *et al*., 2016). The raster was computed using the addresses of the 938 veterinary services registered in the official Swiss Yellow Pages online database in 2013 (Swisscom Ltd., 2016), and the Swiss road network in 2008 was derived from the VECTOR25 data model of the Swiss Federal Office of Topography (SFOT, 2016). Distances to the closest veterinary service were averaged for each municipal unit to provide a measure of average travel distance to veterinary care within a given municipality (Bliss *et al*., 2012). We used more recent information on the addresses of veterinary services and on the areal extent of residential land because data for 2008 is currently not available. Given the information provided by governmental agencies (FOPH, 2016; SFSO, 2016), this was seen as a reasonable compromise for the purpose of this case study.

Filtering out the structural zeros

Sampling is the process of selecting a representative number of individuals to draw inferences about the entire population (Thompson, 2012). In epidemiological research, this process is employed to perform a selection of sampling units, defined as individuals or groups of individuals, allowing an investigation of linkages to disease determinants, for instance, in cohort or case-control studies (Pearce, 2012; Woodward, 2013). Importantly, these stud-

ies are meant to inform about the disease in the at-risk population from which the sampling units have been drawn (Pearce 2012; Woodward, 2013). Various methods can be employed to define sampling units, using random and non-random designs (Cattin, 1980; Banerjee and Chaudhury, 2010). While random sampling is designed to produce generalisable results, non-random sampling is critical because the representativeness for the entire at-risk population is not possible, and therefore the results of the epidemiological study might not be generalisable (Cattin, 1980; Banerjee and Chaudhury, 2010). Although the sampling of enumerated data is exceptional in epidemiological research (Nejjari *et al*., 1993; Lawson, 2006), in this case study, we carried out a non-random selection of the Swiss municipal units where cancer diagnostic examinations have been performed in the year of interest. This filtering step, which discarded all structural zeros, was felt to be justified by the need to draw a representative sample to evaluate the effects of under-ascertainment in models of canine cancer incidence (Cattin, 1980; Banerjee and Chaudhury, 2010). Therefore, in parallel, we also fitted the model for the complete enumeration across all Swiss municipal units, which include structural zeros. We compared the statistical distributions (Oja, 1983) as well as statistical performance and predictive power (Snee, 1977; Picard and Cook, 1984) of the two modelling scenarios to evaluate changes associated with our filtering step, and thus identify direct effects of structural zeros on the model of canine cancer incidence.

Modelling canine cancer incidence

We fitted canine cancer incidence in a Poisson regression framework. However, the incidence data might deviate from a standard Poisson distribution, which occurs when the variance is not equal to the mean of the incidence data (Cameron and Trivedi, 1990; Berk and MacDonald, 2008). We did not test alternative regression frameworks, such as negative binomial (Hardin *et al*., 2007; Berk and MacDonald, 2008), zero-inflated and hurdle (Hu *et al*., 2011; He *et al*., 2014) models, because the coefficients accommodating different statistical distributions impede a direct comparison between the two modelling scenarios (Preisser *et al*., 2012; Arab, 2015). In addition, the relatively simple structure of the Poisson regression framework allows a more straightforward assessment of potential changes in the coefficient estimates for each independent variable (Arab, 2015). We fitted the observed canine cancer incidence (y) through the following independent variables (x): *canine population size*, *canine average age*, *canine female ratio*, *average income tax*, *human population density*, and *distance to veterinary care*. The predicted canine cancer incidence ($\hat{y}$) was log-transformed according to Equation 1, presented below. In Equation 1, θ' denotes α concatenated to β: two parameters of the model that are estimated by maximum likelihood (Frome, 1983; Frome and Checkoway, 1985).

$$\log(\hat{y}(y|x)) = \theta' x \qquad \text{Eq. 1}$$

To contrast the two modelling scenarios, we investigated significance levels (α=.05) and changes in the coefficient estimates, as well as the proportion of variance reduction η^2 for each independent variable (Pearson, 1911; Fisher, 1928). In doing so, we focused on potential changes occurring between *canine population size*, *canine average age* and *canine female ratio*, and *average income tax*, *human population density* and *distance to veterinary care*. These two sets of independent variables inform about two distinct elements: demographic risk factors and potential under-ascertainment. We then computed the McFadden pseudo-R-Squared as a measure of statistical performance of the two modelling scenarios (Cameron and Windmeijer, 1996, 1997) and mapped the Pearson residuals to identify municipal units of poor model predictions as well as potential spatial non-stationarity in the statistical associations (Brunsdon *et al*., 1996; Fotheringham *et al*., 1996). We perused Pearson residuals because these can highlight an important lack of model fit, *i.e.* when the absolute values exceed 2.0, and especially 3.0 (Cameron and Windmeijer, 1996, 1997).

Cross-validating the two modelling scenarios

Given that the incidence data fit in the two modelling scenarios present different statistical distributions (Frome, 1983; Frome and Checkoway, 1985), we employed a cross-validation method based on 1000 model iterations for contrasting statistical performance and predictive power of the models (Snee, 1977; Picard and Cook, 1984). Cross-validation allows assessing how well the model of canine cancer incidence will generalise to a different dataset (Snee, 1977; Picard and Cook, 1984). This is a critical issue for potential comparative studies of canine and human cancers (Kukull and Ganguli, 2012; St. Sauver *et al*., 2012). For each model iteration, we randomly fitted 80% of the municipal units (*i.e*. the training set) to predict the remaining 20% (*i.e*. the validation set) (Snee, 1977; Picard and Cook, 1984). We then assessed central tendency and spread of the coefficient estimates across iterations by means of boxplots (Williamson *et al*., 1989; Gohil, 2015) to evaluate the stability of statistical relationships across iterations, and thus statistical performance (Snee, 1977; Picard and Cook, 1984). We also computed measures of predictive power by averaging the mean absolute error (MAE) and the root mean square error (RMSE) across iterations (Willmott, 1981; Hyndman and Koehler, 2006). The MAE is an absolute measure of the error, defined as the difference between the predicted ($\hat{y}$) and observed (y) canine cancer incidence, as presented in Equation 2 (Willmott, 1981).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \qquad \text{Eq. 2}$$

In addition, to better understand the skewness of the error distribution, we assessed the 50th, 90th and 95th percentiles of the MAE. These measures were also averaged for the two modelling scenarios (Willmott, 1981). We then computed the RMSE, which is defined as the square root of the squared error, as presented in Equation 3 (Willmott, 1981).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad \text{Eq. 3}$$

In computing the average RMSE, we discarded the model iterations resulting in outliers through a standard single-step outlier-detection procedure (Hawkins, 1980; Maimon and Rokach, 2010), because this measure is known to be sensitive to large errors (Chai and Draxler, 2014). We finally assessed associations between RMSE and the independent variables used to fit the training set across model iterations to assess whether the latter drive the variation in the error measures for the two modelling scenarios (Hirsch, 1991). This exploratory step was performed through Spearman's ρ correlation tests (Spearman, 1904), where significant associations (α=.05) were further investigated by means of scatterplots (Gohil, 2015).

Results

Exploring the two modelling scenarios

Figure 1 presents the spatial distribution of the canine cancer incidence retrieved from the Swiss Canine Cancer Registry in 2008, fitted in the two modelling scenarios described earlier. Figure 1A shows the incidence data for the modelling scenario based on the complete enumeration (*i.e.* including all Swiss municipalities), where 1298 municipal units out of 2351 indicate zero incidences. Figure 1B shows the filtered incidence data, where 939 municipal units were identified as having structural zeros and were thus labelled as *no data*. As a consequence of the filtering step, only 359 municipal units out of 1412 exhibited zero incidences. These numbers demonstrate that structural zeros are widespread in the incidence data retrieved from the Swiss Canine Cancer Registry in 2008, and that the statistical distributions, estimated through measures of central tendency and spread, could be substantially different between the two modelling scenarios (Oja, 1983). Indeed, the mean and median of the incidence data were 1.5 and 0.0 for the complete enumeration, and to 2.5 and 1.0 for the filtered sample. The coefficient of variation of the incidence data was 334% for the complete enumeration, and 250% for the filtered sample. These measures show that, for the two modelling scenarios, the incidence data deviated from a standard Poisson distribution (Cameron and Trivedi, 1990; Berk and MacDonald, 2008), and this deviation was greater in the complete enumeration because of the higher coefficient of variation. In addition, Figure 1B suggests that zeros were not randomly distributed across the study area. Regions of municipal units with sampling zeros (*i.e.* zero incidences) and structural zeros (*i.e. no data*) could be identified in the Alps and the Jura Mountains. Based on prior research (Boo *et al*., 2015, 2016), these rural regions were expected to show higher degrees of under-ascertainment as this is associated with a lower urban character and greater travel distances to veterinary care. When fitting the incidence data, all coefficient estimates presented in Table 1 were statistically significant (α=.05) and remained very similar across the two modelling scenarios. Table 1 shows that *canine average age* produced negative coefficient estimates. This statistical association contradicts existing findings, indicating a higher prevalence of cancer in older dogs, thus suggesting that under-ascertainment might be critical for this age class (Bronson, 1982; Eichelberg and Seine, 1996; Lund *et al*., 1999; Michell, 1999; Proschowsky *et al*., 2003). Conversely, *canine population size* and *canine female ratio* both produced positive coefficient estimates. Such statistical associations imply a higher incidence of cancer in larger at-risk canine populations and a higher prevalence of cancer in female dogs (Bronson, 1982; Eichelberg and Seine, 1996; Lund *et al*., 1999; Michell, 1999; Proschowsky *et al*., 2003). *Average income tax* and *human population density* both produced positive coefficient estimates, thus confirming that the ascertainment of cancer might improve in municipalities with higher socioeconomic status and urban lifestyle (Boo *et al*., 2015, 2016). Finally, *distance to veterinary care* produced negative coefficient estimates, indicating that under-ascertainment of cancers can be linked to greater travel distance to veterinary care (Boo *et al*., 2016). Table 1 also indicates that, altogether, the independent variables accounting for potential under-ascertainment (*i.e. average income tax*, *human population density* and *distance to veterinary care*) have a lower proportion of variance reduction for the filtered sample (η^2=0.36) compared to the complete enumeration (η^2=0.46). Such an increase could be linked to higher degrees of under-ascertainment in the latter modelling scenario. When assessing the statistical performance for the two modelling scenarios, the McFadden pseudo-R-squared measures indicated a slight increase in statistical performance for the model based on the filtered sample (R^2=0.32), compared to the complete enumeration (R^2=0.31) (Cameron and Windmeijer, 1996, 1997). As previously mentioned, such improvement might be linked to the change of the statistical distribution of the incidence data used to fit the two modelling scenarios (Cameron and Trivedi, 1990; Berk and MacDonald, 2008).

Figure 2 presents the spatial distribution of Pearson residuals for the two modelling scenarios to explore model fit across the study area (Cameron and Windmeijer, 1996, 1997). Figure 2A shows that for the complete enumeration most municipal units located in the Alps and the Jura Mountains were characterised by acceptable model over-estimations, with residuals between −1.9 and −0.1. This result can be linked to the large part of zero incidences (*i.e.* both structural and sampling zeros) in these rural regions, which are captured by the model of canine cancer incidence. Figure 2B illustrates an increased predictive power for the filtered sample. This is because most municipal units with residuals above 2.0 and below −2.0 in the complete enumeration showed residuals between −1.9 and 1.9 in the filtered sample. The two modelling scenarios also exhibited several regions with residuals above 3.0, which are typically located within urban agglomerations, for instance, in Zurich, Basel or Berne. In these urban regions, residuals could reach 22.8 for the complete enumeration and 17.7 for the filtered sample. On the one hand, these results suggest local spatial autocorrelation of model residuals and thus potential spatial non-stationarity of statistical associations

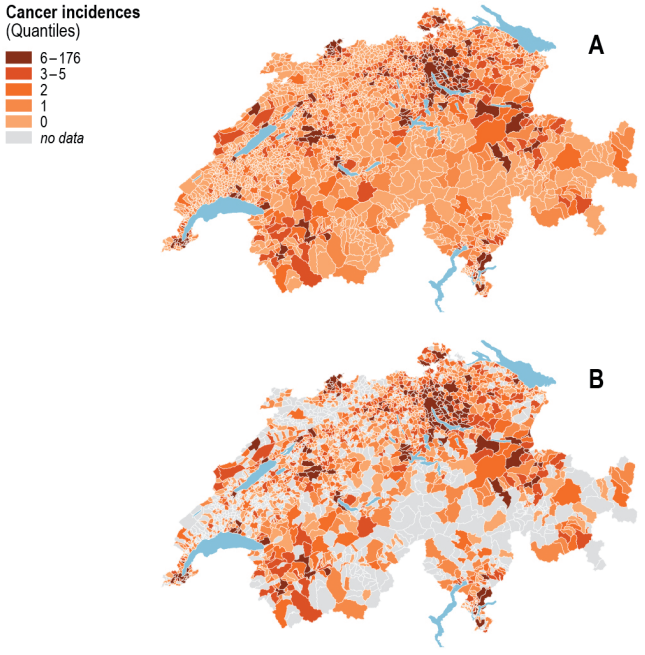

Figure 1. Spatial distributions of the cancer incidences across the two modelling scenarios: A) complete enumeration; B) filtered sample.

(Brunsdon *et al*. 1996; Fotheringham *et al*. 1996). On the other hand, the differences between the two modelling scenarios indicate that the filtered sample results in a better model fit for most municipal units (Cameron and Windmeijer, 1996, 1997). However, such improvements might not be detected when computing averaged measures of predictive power, such as MAE and RMSE, because the model based on the complete enumeration presented a larger number of municipal units with residuals between −1.9 and 1.9 (Willmott, 1981; Hyndman and Koehler, 2006).

Comparing the two modelling scenarios through cross-validation

We further compared statistical performance and predictive power for the two modelling scenarios through model cross-validation, which was based on a training-/validation-set ratio of 1881/470 municipal units for the complete enumeration, and 1130/282 municipal units for the filtered sample. The boxplots in Figure 3 depict the spread of the coefficient estimates over the 1000 model iterations for the two modelling scenarios. The median is shown as a thick black line, the interquartile range is indicated as a grey box and the minimum and maximum estimates are delimited by the whiskers (Williamson *et al*., 1989; Gohil, 2015). In Figure 3, the median values for the two modelling scenarios looked very similar to the coefficient estimates presented in Table 1, thus suggesting overall stability across iterations (Snee, 1977; Picard and Cook, 1984). Still, when comparing the distribution of the coefficient estimates across the two modelling scenarios, *canine population size*, *average income tax*, *human population density* and *distance to veterinary care* showed a decreased spread for the filtered sample. This indicates an increased stability of coefficient estimates across iterations for the independent variables accounting for the size of the at-risk population and for potential under-ascertainment, thus suggesting improved statistical performance for the filtered sample (Snee, 1977; Picard and Cook, 1984). Independent variables such as *canine average age* and *canine female ratio* showed similar spreads for the two modelling scenarios, possibly because the large portion of zeros in the complete enumeration stabilised these coefficients towards zero. Despite the similar spread, *canine female ratio* showed a change of sign in some of the model iterations for the complete enumeration. Such a statistical association contradicts prior research (Bronson, 1982; Eichelberg and Seine, 1996; Lund *et al*., 1999; Michell, 1999; Proschowsky *et al*., 2003) and further confirms that the complete enumeration could produce biased coefficient estimates that might, in turn, result in less accurate predictions of canine cancer incidence (Snee, 1977; Picard and Cook, 1984).

Table 2 shows that the average MAE for the complete enumeration was four times larger than for the filtered sample, thus indicating a substantial improvement of predictive power for the latter modelling scenario (Willmott, 1981; Hyndman and Koehler, 2006). The averaged percentiles of the MAE showed that the average error distribution was heavily skewed in the complete enumeration as, on average, only 5% of the errors (*i.e*. above the 95th percentile) were accountable for a higher MAE. For the filtered sample, the error distribution appeared less skewed because of a general decrease of the error magnitudes (Willmott, 1981). This skewed error distribution also affected the calculation of the average RMSE for the complete enumeration (Chai and Draxler, 2014). The outlier detection procedure revealed six iterations with RMSE between 3700 and 159,300 for the complete enumeration, and three iterations with RMSE between 5700 and 9000 for the filtered sample. These RMSE outliers were at least 10 times higher than the average RMSE for the two modelling scenarios and were thus removed (Hawkins, 1980; Maimon and Rokach, 2010). After discarding these outliers, the average RMSEs became more meaningful for comparing the two modelling scenarios and still confirm a higher predictive power for the filtered sample (Willmott, 1981; Hyndman and Koehler, 2006).

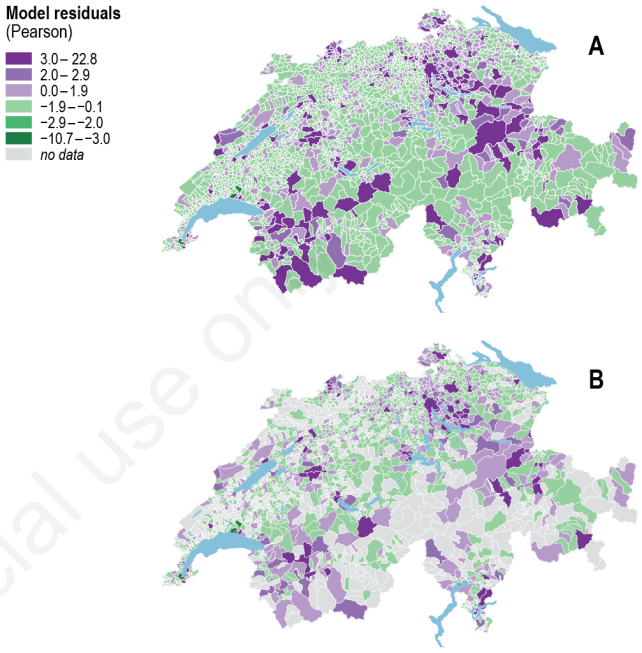

Figure 2. Spatial distributions of the Pearson residuals across the two modelling scenarios: A) complete enumeration; B) filtered sample.

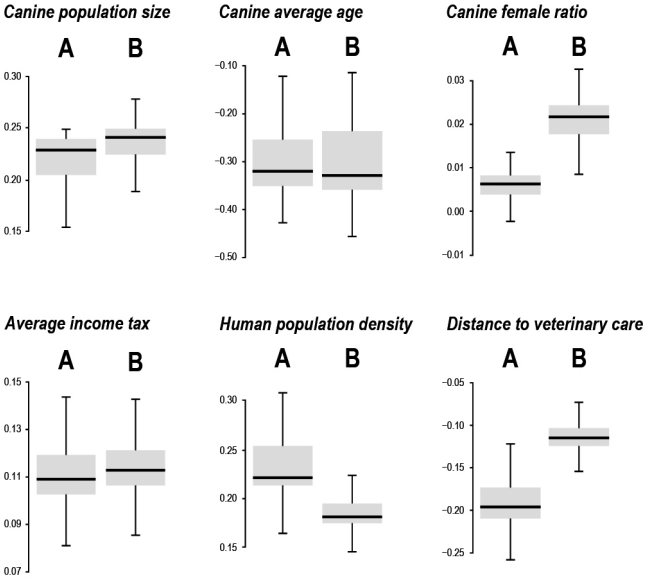

Figure 3. Distributions of the coefficient estimates across the two modelling scenarios: A) complete enumeration; B) filtered sample.

We finally computed Spearman's ρ correlations across the model iterations between the RMSE and the independent variables used to fit the training sets to better understand the role of different variables as drivers of variation in the RMSE measures (Hirsch, 1991). The only significant correlation with the RMSE was found for the average *canine population size*, with a ρ of −0.72 (P=.00) for the complete enumeration and a ρ of −0.76 (P=.00) for the filtered sample. Such strong negative correlation between average *canine population size* used in the training set and the RMSE for the two modelling scenarios is a known association between the size of the sampled population and predictive power, as presented in Figure 4 (Cattin, 1980; Banerjee and Chaudhury, 2010). In the scatterplots, the trend of the correlation is highlighted by the grey surface representing the conditional mean smoothed through a linear model fit (Gohil, 2015). Again, the trend surfaces showed that smaller average canine population sizes in the training sets generally resulted in higher RMSE, and thus in lower predictive power (Cattin, 1980; Banerjee and Chaudhury, 2010). Still, for both modelling scenarios, the RMSE seemed to be partly independent of the average canine population sizes used in the training set, as there are four distinct RMSE clusters with values around 0.0, 5.0, 50.0 and 150.0. These clusters suggest two major issues that might affect the two modelling scenarios. First, it appeared that there is a need to include additional, unspecified independent variables in the two modelling scenarios (Hirsch, 1991; Allen, 2007). This is not surprising considering that the independent variables included in the model of canine cancer incidence only account for demographic risk factor and potential under-ascertainment (Boo *et al*., 2015, 2016). Second, these clusters further corroborated the hypothesis of potential spatial non-stationarity in the statistical associations derived from the spatial distributions of model residuals presented in Figure 2 (Brunsdon *et al*., 1996; Fotheringham *et al*., 1996).

Discussion

By systematically comparing the two modelling scenarios we were able to provide insights into effects of structural zeros on models of canine cancer incidence. The exceptionally rich attribution of the Swiss Canine Cancer Registry allowed us to uncover a large number of structural zeros in the incidence data retrieved within the Swiss municipal units in 2008 (Hu *et al*., 2011; He *et al*., 2014). Structural zeros can affect the statistical distributions of the incidence data, as indicated by the measures of central tendency and spread, which illustrated a deviation from the Poisson distribution (Cameron and Trivedi, 1990; Berk and MacDonald, 2008). For this reason, we expected that the presence of structural zeros would also impact the statistical performance of the Poisson regression model (Frome, 1983; Frome and Checkoway, 1985). However, we were not able to detect a substantial change in the McFadden pseudo-R-squared measures (Cameron and Windmeijer, 1996, 1997), and the same is true for the significance level of the coefficient estimates (Frome, 1983; Frome and Checkoway, 1985). Furthermore, we were unable to identify any substantial changes in the coefficient estimates, except for a higher proportion of variance reduction η^2 for the independent variables accounting for potential under-ascertainment (Pearson, 1911; Fisher, 1928). The latter suggests that the model of canine cancer

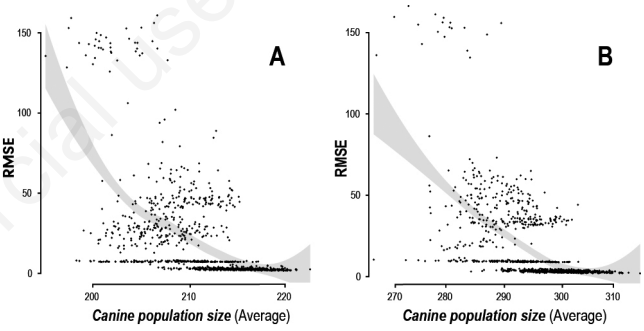

Figure 4. Canine population size *vs* root mean square error across the two modelling scenarios: A) complete enumeration; B) filtered sample.

Table 1. Coefficient estimates, P values and proportions of variance reduction η2 across the two modelling scenarios.

	Complete enumeration			Filtered sample		
	Estimate	P	η^2	Estimate	P	η^2
Canine population size	0.23	.00	0.54	0.24	.00	0.62
Canine average age	−0.32	.00	0.00	−0.33	.00	0.01
Canine female ratio	0.01	.03	0.00	0.02	.00	0.01
Average income tax	0.11	.00	0.07	0.12	.00	0.08
Human population density	0.23	.00	0.28	0.19	.00	0.23
Distance to veterinary care	−0.20	.00	0.11	−0.11	.00	0.05

Table 2. Mean absolute error percentiles and root mean square error averaged across the two modelling scenarios.

	Average MAE				Average RMSE	
	Raw	50th	90th	95th	Raw	Outliers removed
Complete enumeration	19.35	0.90	2.58	4.29	362.15	20.25
Filtered sample	4.29	1.26	3.32	4.98	40.88	17.28

MAE, mean absolute error; RMSE, root mean square error.

incidence might capture structural zeros as a manifestation of a potential under-ascertainment. This is further supported by the spatial distribution of the observed cancer incidence, as presented in Figure 1, which shows that most structural zeros occurred across municipal units located in rural regions, where we expected potential under-ascertainment to be more important (Boo *et al*., 2015, 2016). The spatial distribution of Pearson residuals presented in Figure 2 illustrates a general decrease of predictive power for a number of municipal units, which indicates critical model over- and under-estimation (Willmott, 1981; Hyndman and Koehler, 2006). However, the municipal units presenting structural zeros result, in most cases, in acceptable model over-estimations (Willmott, 1981; Hyndman and Koehler, 2006). This is because the predicted canine cancer incidence is also very close to zero. As a consequence, the number of municipal units with an acceptable model over-estimation becomes prominent, and the impact of the critical model over- and under-estimations might thus be difficult to detect (Willmott, 1981; Hyndman and Koehler, 2006).

Through model cross-validation, we were able to detect an increased spread of the coefficient estimates for nearly all independent variables, and for the independent variable *canine female ratio,* we could even observe changes of sign in the coefficient estimates. These findings indicate that the presence of structural zeros is critical for the statistical performance of the model of canine cancer incidence because statistical relationships become less stable across model iterations (Snee, 1977; Picard and Cook, 1984). We also observed an average MAE that is four times larger if structural zeros are involved, a decrease in predictive power related to less than 5% of the errors (Willmott, 1981). These findings, together with similar observations for the average RMSE, suggest that structural zeros severely affect the predictive power of the model of canine cancer incidence (Willmott, 1981; Hyndman and Koehler, 2006). Yet, we were unable to detect significant Spearman's ρ correlations between the selected independent variables and the RMSE across model iterations that could be linked to the sole presence of structural zeros (Hirsch, 1991). The results of the model cross-validation allowed us to discover important effects of structural zeros on the model of canine cancer incidence and how they are critical for the generalisation to a different dataset (Snee, 1977; Picard and Cook, 1984). This is a crucial issue because it challenges future comparative studies of canine and human cancers based on the incidence data retrieved from the Swiss Canine Cancer Registry (Snee, 1977; Picard and Cook, 1984).

Conclusions

This case study shows that epidemiological studies could greatly benefit from increased awareness about the presence of structural zeros in the incidence data retrieved from canine cancer registries. New insights into effects of structural zeros on models of canine cancer incidence have been provided. Structural zeros occur when zero incidences originate from the sole absence of performed diagnostic examinations within a given sample unit. These structural phenomena are particularly critical as they are often mistaken for sampling zeros that, in turn, result from diagnostic examinations performed within a given sample unit. The exceptionally rich attribution of the Swiss Canine Cancer Registry enabled us to identify a large number of structural zeros in the incidence data retrieved across Swiss municipalities in 2008. We investigated effects of structural zeros on models of canine cancer incidence based on independent variables accounting for demographic risk factors and for potential under-ascertainment (Boo *et al*., 2015, 2016). The results of our modelling effort show that it is rather difficult to identify effects of structural zeros through simple goodness-of-fit and significance tests (Cameron and Windmeijer, 1996, 1997). In this regard, the cross-validation framework represented an effective approach for detecting the impacts related to the presence of structural zeros. We identified increased instability in the statistical associations (Snee, 1977; Picard and Cook, 1984) as well as increased error measures, (Willmott, 1981; Hyndman and Koehler, 2006). These results indicate an overall decrease of statistical performance and predictive power, and they suggest that the presence of structural zeros challenges the generalisability of the model of canine cancer incidence to a different dataset (Snee, 1977; Picard and Cook, 1984). A better understanding of these effects is thus critical for modelling canine cancer incidence, as well as for future comparative studies of canine and human cancers based on incidence data retrieved from the Swiss Canine Cancer Registry (Kukull and Ganguli, 2012). We thus contend that these results should be considered as a first step to develop further investigations into the effects of structural zeros in epidemiological research.

The results of this case study also raise new questions that will drive our future epidemiological studies of canine cancer incidence in Switzerland. First, there is a clear need for model specification, which involves the inclusion of additional independent variables in the model of canine cancer incidence (Hirsch, 1991), as well as accounting for deviations from the Poisson distribution (Cameron and Trivedi, 1990; Berk and MacDonald, 2008). For this reason, in future studies, we will model incidence data for canine cancers that have strong linkages to environmental determinants (Schmidt, 2009; Reif, 2011), and we will test for possible over-dispersion and zero-inflation in the incidence data (Preisser *et al*., 2012; Arab, 2015). Another interesting question is connected with the spatial autocorrelation of model residuals, which suggests spatial non-stationarity of statistical associations (Brunsdon *et al*., 1996; Fotheringham *et al*., 1996). As this condition implies that fitting the model for the whole study area cannot explain the associations with the selected independent variables (Brunsdon *et al*., 1996; Fotheringham *et al*., 1996), we will test different strategies for modelling the incidence data assuming local and regional variations in the spatial structure (Assunção, 2003; Chen *et al*., 2015).

References

Allen MP, 2007. Model specification in regression analysis. In: M.P. Allen (ed.) Understanding regression analysis. Springer, Berlin, Germany, pp. 166-170.

ANIS, 2016. Animal identity service AG. Available from: http://www.anis.ch

Arab A, 2015. Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. Int J Environ Res Pub Health 12:10536-48.

Assunção RM, 2003. Space varying coefficient models for small area data. Environmetrics 14:453-73.

Banerjee A, Chaudhury S, 2010. Statistics without tears: populations and samples. Ind Psych J 19:60-5.

Berk R, MacDonald JM, 2008. Overdispersion and Poisson regression. J Quant Criminol 24:269-84.

Bliss RL, Katz JN, Wright EA, Losina E, 2012. Estimating proximity to care: are straight line and zipcode centroid distances acceptable proxy measures? Med Care 50:99-106.
Boo G, Fabrikant SI, Leyk S, 2015. A novel approach to veterinary spatial epidemiology: dasymetric refinement of the Swiss Dog Tumor Registry data. ISPRS Annals II-3/W5:263–269.
Boo G, Leyk S, Fabrikant SI, Pospischil A, 2016. A regional approach for modeling dog cancer incidences with regard to different reporting practices. In: Miller J, O'Sullivan D and Wiegand N. (eds.) Geographic Information Science 9th International Conference, GIScience 2016, Montreal, QC, Canada, September 27-30, 2016, Proceedings. Springer, Berlin, Germany, pp. 29-32.
Brønden LB, Flagstad A, Kristensen AT, 2007. Veterinary cancer registries in companion animal cancer: a review. Vet Comp Oncol 5:133-44.
Bronson RT, 1982. Variation in age at death of dogs of different sexes and breeds. Am J Vet Res 43:2057-9.
Brunsdon C, Fotheringham AS, Charlton M, 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr Anal 28:281-98.
Bukowski JA, Wartenberg D, Goldschmidt M, 1998. Environmental causes for sinonasal cancers in pet dogs, and their usefulness as sentinels of indoor cancer risk. J Toxicol Environ Health 54:579-91.
Cameron AC, Trivedi PK, 1990. Regression-based tests for overdispersion in the Poisson models. J Econometrics 46:347-64.
Cameron AC, Windmeijer FAG, 1996. R-Squared measures for count data regression models with applications to health-care utilization. J Business Econ Stat 14:209-20.
Cameron AC, Windmeijer FAG, 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. J Econometrics 77:329-42.
Cattin P, 1980. Estimation of the predictive power of a regression model. J Appl Psychol 65:407-14.
Chai T, Draxler RR, 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247-50.
Chen S, Jiang C, Waller L, 2015. Automatic region-wise spatially varying coefficient regression model: an application to national cardiovascular disease mortality and air pollution association study. Available from: https://arxiv.org/pdf/1511.05924
Delamater PL, Messina JP, Shortridge AM, Grady SC, 2012. Measuring geographic access to health care, raster and network-based methods. Int J Health Geogr 11:15.
Dvorzak M, Wagner H, 2016. Sparse Bayesian modelling of underreported count data. Stat Model 16:24-46.
Eichelberg H, Seine R, 1996. Life expectancy and cause of death in dogs. The situation in mixed breeds and various dog breeds. Berliner Münchener Tierärztliche Wochenschrift 109:292-303.
Fisher RA, 1928. Statistical methods for research workers. Oliver and Boyd, Edinburgh, UK.
FOPH, 2016. Federal office of public health: MedReg. Available from: http://www.medregom.admin.ch
Fotheringham AS, Charlton M, Brunsdon C, 1996. The geography of parameter space: an investigation of spatial non-stationarity. Geogr Inf Syst 10:605-27.
Frome EL, 1983. The analysis of rates using poisson regression models. Biometrics 39:665-74.
Frome EL, Checkoway H, 1985. Use of Poisson regression models in estimating incidence rates and ratios. Am J Epidemiol 121:309-23.
Gavazza A, Presciuttini S, Barale R, Lubas G, Gugliucci B, 2001. Association between canine malignant lymphoma, living in industrial areas, and use of chemicals by dog owners. J Vet Int Med 15:190-5.
Gibbons CL, Mangen M-JJ, Plass D, Havelaar AH, Brooke R, Kramarz P, Peterson KL, Stuurman AL, Cassini A, Fèvre EM, Kretzschmar MEE, 2014. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health 14:147.
Glickman LT, Domanski LM, Maguire TG, Dubielzig RR, Churg A, 1983. Mesothelioma in pet dogs associated with exposure of their owners to asbestos. Environ Res 32:305-13.
Glickman LT, Schofer FS, McKee LJ, Reif JS, Goldschmidt MH, 1989. Epidemiologic study of insecticide exposures, obesity, and risk of bladder cancer in household dogs. J Toxicol Environ Health 28:407-14.
Gohil A, 2015. R data visualization coockbook. Packt Publishing, Birmingham, UK.
Grüntzig K, Graf R, Boo G, Guscetti F, Hässig M, Axhausen KW, Fabrikant SI, Welle M, Meier D, Folkers G, Pospischil A, 2016. Swiss Canine Cancer Registry 1955-2008: occurrence of the most common tumour diagnoses and influence of age, breed, body size, sex and neutering status on tumour development. J Comp Pathol 155:156-70.
Grüntzig K, Graf R, Hässig M, Welle M, Meier D, Lott G, Erni D, Schenker NS, Guscetti F, Boo G, Axhausen KW, Fabrikant SI, Folkers G, Pospischil A, 2015. The Swiss Canine Cancer Registry: a retrospective study on the occurrence of tumours in dogs in Switzerland from 1955 to 2008. J Comp Pathol 152:161-71.
Hardin JW, Hilbe JM, Hilbe J, 2007. Generalized linear models and extensions. Stata Press, College Station, TX, USA.
Hawkins DM, 1980. Identification of outliers. Springer, Berlin, Germany.
Hayes HM, Hoover R, Tarone RE, 1981. Bladder cancer in pet dogs: a sentinel for environmental cancer? Am J Epidemiol 114:229-33.
He H, Tang W, Wang W, Crits-Christoph P, 2014. Structural zeroes and zero-inflated models. Shanghai Arch Psychiatry 26:236-42.
Hirsch RP, 1991. Validation samples. Biometrics 47:1193-4.
Hu M-C, Pavlicova M, Nunes EV, 2011. Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am J Drug Alcohol Abuse 37:367-75.
Hyndman RJ, Koehler AB, 2006. Measuring forecast accuracy. Int J Forecasting 22:679-88.
Kukull WA, Ganguli M, 2012. Generalizability. Neurology 78:1886-91.
Lawson AB, 2006. Statistical methods in spatial epidemiology. John Wiley & Sons, Chichester, UK.
Legendre P, Legendre LFJ, 2012. Numerical ecology. Elsevier, Amsterdam, The Netherlands.
Lund EM, Armstrong PJ, Kirk CA, Kolar LM, Klausnor JS, 1999. Health status and population characteristics of dogs and cats examined at private veterinary practices in the United States. Am Vet Med Assoc 214:1336-41.
Maimon O, Rokach L, 2010. Data mining and knowledge discovery handbook. Springer, Berlin, Germany.
Michell AR, 1999. Longevity of British breeds of dog and its rela-

tionships with sex, size, cardiovascular variables and disease. Vet Rec 145:625-9.
Mohri M, Roark B, 2005. Structural zeros versus sampling zeros. Oregon Health & Science University, Portland, OR, USA.
Nejjari C, Tessier JF, Dartigues JF, Barberger-Gateau P, Letenneur L, Salamon R, 1993. The relationship between dyspnoea and main lifetime occupation in the elderly. Int J Epidemiol 22:848-54.
Nødtvedt A, Berke O, Bonnett BN, Brønden L, 2012. Current status of canine cancer registration – report from an international workshop. Vet Comp Oncol 10:95.101.
Oja H, 1983. Descriptive statistics for multivariate distributions. Stat Prob Lett 1:327-32.
Owen LN, 1979. A comparative study of canine and human breast cancer. Invest Cell Pathol 2:257-75.
Pearce N, 2012. Classification of epidemiological study designs. Int J Epidemiol 41:393-7.
Pearson K, 1911. On a correction needful in the case of the correlation ratio. Biometrika 8:254-6.
Picard RR, Cook RD, 1984. Cross-validation of regression models. J Am Stat Assoc 79:575-83.
Pinho SS, Carvalho S, Cabral J, Reis CA, Gärtner F, 2012. Canine tumors: a spontaneous animal model of human carcinogenesis. Transl Res 159:165-72.
Pospischil A, Hässig M, Vogel R, Salvini MM, Fabrikant SI, Axhausen K, Schenker SN, Erni D, Guscetti F, 2013. Hundepopulation und Hunderassen in der Schweiz von 1955 bis 2008. Schweizer Archiv für Tierheilkunde 155:219-28.
Preisser JS, Stamm JW, Long DL, Kincade ME, 2012. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Res 46:413-23.
Proschowsky HF, Rugbjerg H, Ersbøll AK, 2003. Mortality of purebred and mixed-breed dogs in Denmark. Prev Vet Med 58:63-74.
Reif JS, 2011. Animal sentinels for environmental and public health. Publ Health Rep 126:50-7.
Reif JS, Bruns C, Lower KS, 1998. Cancer of the nasal cavity and paranasal sinuses and exposure to environmental tobacco smoke in pet dogs. Am J Epidemiol 147:488-92.
Schmidt PL, 2009. Companion animals as sentinels for public health. Vet Clin N Am 39:241-50.
Scotch M, Odofin L, Rabinowitz P, 2009. Linkages between animal and human health sentinel data. BMC Vet Res 5:15.
SFOT, 2016. Federal Office of Topography-Swisstopo. Available from: http://www.swisstopo.admin.ch
SFSO, 2016. Swiss Federal Statistical Office. Available from: http://www.bfs.admin.ch
SFTA, 2016. Swiss Federal Tax Administration. Available from: http://www.estv.admin.ch
Snee RD, 1977. Validation of regression models: methods and examples. Technometrics 19:415-28.
Spearman C, 1904. The proof and measurement of association between two things. Am J Psychol 15:72-101.
St. Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ, Rocca WA, 2012. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester epidemiology project. Mayo Clin Proc 87:151-60.
Swisscom Ltd, 2016. The official phonebook and yellow pages of Switzerland. Available from: http://www.local.ch
Thompson SK, 2012. Sampling. Wiley, Hoboken, NJ, USA.
Vail DM, MacEwen EG, 2000. Spontaneously occurring tumors of companion animals as models for human cancer. Cancer Invest 18:781-92.
Vascellari M, Capello K, Carminato A, Zanardello C, Baioni E, Mutinelli F, 2016. Incidence of mammary tumors in the canine population living in the Veneto Region (Northeastern Italy): Risk factors and similarities to human breast cancer. Prev Vet Med 126:183-9.
Williamson DF, Parker RA, Kendrick JS, 1989. The box plot: a simple visual method to interpret data. Ann Intern Med 110:916-21.
Willmott CJ, 1981. On the validation of models. Phys Geogr 2:184-94.
Woodward M, 2013. Epidemiology: study design and data analysis. CRC Press, New York, NY, USA.