Estimation of Echinococcosis Cyst Prevalence in Two Rural Communities Considering spatial distribution in the selection bias

O G E Espinoza-Hurtado[1*], E A Chacón-Montalvan[2], Saúl Santivañez[1,3]

1. Global Health Center, Cayetano Heredia Peruvian University, Lima, Peru
2. Mathematics and Statistics Department, Lancaster University, Lancaster, United Kingdom
3. National University of the Center of Peru, Junin, Peru

* oespinozah@uni.pe

Abstract

Objective: Determine a correction factor to correctly estimate the prevalence of hydatidosis. **Methods.** Prevalence was estimated by weighted likelihood. Each point was considered as a realization of a Poisson process (marked for the case of the sample). Intensity functions were estimated as Poisson process. Risk function was estimated using a GAM Weighted of binomial family and logistic transformation as link function. La correct value of ρ was the OR of the disease between the sample and the population. **Results and Discussion.** The correction factor was 1.216 and with it there was a fixed prevalence of 0.207. The intensity graphs show that the overestimation could have been due to an overcollection of cases at certain points. Georeferencing could become an alternative to avoid overestimating without expanding coverage sampling.
Keyword: *Hydatidosis prevalence, Spatial distribution, Sample selection bias correction, Re-weighting.*

Introduction

Correctly estimate epidemiological indicators of a disease is of utmost importance for a country. This because of the impact generated by a zoonosis on the economy of a society consists not only of the money invested for its treatment and prevention or the losses that it causes in livestock and agricultural activities; but also includes disability and lifestyle change that it and his treatment entails [1]. Some diseases, for example hydatidosis [2], are endemic in certain regions. Therefore, they are considered as a public health problem and the underestimation or overestimation of their effects and scope have an impact on their control [3].

In prospective cohort studies conducted in a community there is often a selection bias (bias given by a systematic error in the collection [4]). When sampling is done for convenience, the ratio of cases and controls could be different between the sample and the population. As a result, the traditional way to estimating prevalence as a proportion [5], where n_1 and n_2 are the number of cases and controls in the sample, would determine an incorrect value for it.

$$\hat{\theta} = \frac{n_1}{n_1 + n_2} \tag{1}$$

The over-collection of cases in the sample may be due to their predisposition to participate in the study. But it can also be given by the location of the health center where the

information was collected. In the case when the bias effect is significant, it is necessary to integrate a correction factor ρ in the estimation based on the Weighted Likelihood Estimation for a Bernoulli distribution.

$$\tilde{\theta} = \frac{n_1}{n_1 + \rho\, n_2} \tag{2}$$

Several studies have tried to fix the bias under other approaches and for other purposes. A study on hydatidosis have used the propensity score as a tool to fix bias in order to estimate OR; but not the prevalence [6]. In contrast, a study carried out in China [7] considered the heterogeneity of the probability that each individual has to be selected for a sample in estimating the number of cancer patients, treating the information obtained under a Bayesian approach for a case of recapture and recapture. The limitation of this study is the need to have a very large area and more than one health center in it. By other hand, considering the spatial distribution of the individuals captured in a given sample has been used in the field of ecology [8], which can be replicated in the field of epidemiological sciences [9].

With the goal of estimating hidatidosis prevalence in Corpacancha we propose a spatial marked point processes with additive effects that is able to take into acount the trends. For that, was necessary determine the correction of case-control studies under bias sampling proposing a spatial marked point process model. In addition, this method could be used in other studies with similar characteristics in its collection.

In materials and methods section, the population was presented and the information gathering process was detailed, and the statistical analysis used is described. Also, the usefulness of the method for other studies is mentioned. In results and discussion section the analysis results are displayed together with graphs in order to compared with traditional methods

Materials and methods

Population under study

The population under the study is composed of the 332 (141 at sample) citizens of Corpacancha, Junin - Peru. For this case, the study did not consider within the population those who live in the slaughterhouse because they are very far from the health center.

Design and sampling

It was a prospective cohort study that began in October 2017. It had 2 different ways, one after the other, to obtain the baseline information from the citizens. The first one consisted of a free health campaign to diagnose and rule out hydatidosis in the village health center. In which to determine the presence of hydatidosis in each patient, an abdominal ultrasound (according to WHO criteria [10]) and a western blot serological test were performed [11]. The second consisted in searching for positive cases directly in the community. In that, georeferencing of each household was made by a census conducted a few months after the health campaign. In this study, the outcome only measured the presence of hydatidosis and did not make a difference between the affected organ (liver or lung). Some missing data respect aged (less than 4 observation) have been inputted using mean.

Statistical analyses

First, the statistical analysis began describing the campaign information and covariates to deteterminate which on has an effect in the prevalence and how work with their values.

Then, taking into consideration that each point ($x_i = \{x_{i1}; x_{i2}\} \in R^2$) have been considered as a realization of a Poisson process [12], it began with an exploratory spatial analysis of the data (intensity with contours) to show the proportion between the sample intensity λ and the population intensity λ_p. Then, it adjusts the intensity function of the sample by

$$\begin{aligned} \lambda(x_i) &= \lambda_p(x_i)h(x_i) \\ &= \lambda_p(x_i)exp\left(\beta_0 + h'(x_i)\right) \\ \log\lambda(x_i) &= \log\lambda_p(x_i) + \beta_0 + h'(x_i) \end{aligned}$$

Using a generalized additive model (GAM), where $\log\lambda_p(x_i)$ is an offset, β_0 is a constant and $h(x_i)$ is the sampling effort por each x_i point. If $h'(x_i) = 0$, $\log\beta_0$ is the sampling proportion. Subsequently, the sample risk function $\pi(x_i)$ was estimated using a GAM Weighted of binomial family and logistic transformation as link function [13] with $\lambda_p(x_i)/\lambda(x_i)$ as weight for each x_i. For this model, it considered both the special effect and other covariates that may explain the presence of the disease optimizing the model's auc. With this rick function, disease was predicted for each unsampled individual. This prediction of unsampled cases and controls is because the optimal value of ρ is given when it is equal to the OR of the disease between the sample and the population.

$$\rho = OR_{sample,population}$$

Finally, the value of ρ was used in Eq 2 to estimate unbiased prevalence.

Method application

The importance of using this method lies in the time space between the two data collection. The corrected value of the prevalence using the data from the campaign and the census, together with the other processed information, can be used as a reference indicator when the principal investigator need to determinate the budget about the second stage of collection.

Results and Discussion

The health campaign obtained a coverage of 42.5%, showing a prevalence of hydatidosis of 0.241 ($IC_{95\%}$ $[0.189; 0.293]$). As Table 1 shows, there is a bit observable difference from disease in covariates.

The coverage could be spatially presented as the proportion between the sample intensity function and the population intensity function. As show Fig 1, coverage is not constant in the space.

The sampling intensity and the sampling effort (Eq. **??**) were determinate by a GAM with 49 base functions. It made possible to fit a GAM about risk considering covariate showed at Table 1. The model was chosen taking as criteria the AUC and Deviance Explained. With 23 base functions for the spatial effect considering sex, 11 base functions for the age considering sex and 5 base functions for the number of dog that each person have, the model had an AUC greater than 0.99. It is important to note that if this model to predict the cases and controls had not considered the spatial effect, nor the reweighting, the AUC of the model would have been 0.80. With this model, the

Covariate		**Positive**	**Negative**
Sex			
	Male	16 (27.6%)	42 (72.4%)
	Female	18 (21.7%)	65 (78.3%)
Age*		36.1	32.4
Number of dogs			
	0	21 (22.1%)	74 (77.9%)
	1	2 (13.3%)	13 (86.7%)
	2	3 (17.6%)	14 (82.4%)
	3	7 (63.6%)	4 (36.4%)
	4	1 (33.3%)	2 (66.7%)

* mean by result

Table 1. Covariates

Model	Train			Test		
	Spef.	Sens.	AUC	Spef.	Sens.	AUC
Age	1	0.0833	0.5578	0.03125	1	0.5672
Age*	0.9333	0.2917	0.6122	0.59375	0.6	0.6453
Sex	1	0	0.5358	1	0	0.5469
Sex*	1	0	0.4642	1	0	0.5469
Dogs	0.9333	0.3333	0.6333	0.9688	0	0.149
Dogs*	0.9333	0.3333	0.6333	0.9688	0	0.149
Age, sex	1	0.0833	0.6369	0.03125	1	0.5406
Age, sex*	0.9467	0.4583	0.8708	0.5625	0.8	0.7562
Age, sex, dogs	0.96	0.375	0.74	0.0938	0.9	0.5141
Age, sex, dogs*	0.9467	0.4583	0.7144	0.625	0.5	0.5328
Full	0.96	0.333	0.7403	0.53125	0.5	0.6
Full*	0.9733	0.375	0.7503	0.5625	0.8	0.7438

Table 2. Model Fit Metrics. *Model fitted with weights.

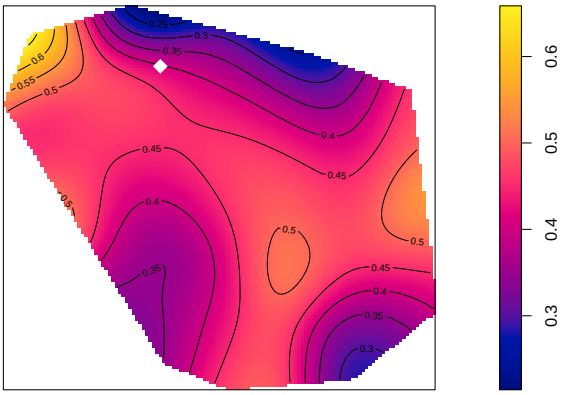

Fig 1. Sampling effort. Proportion between the sample intensity function and the population intensity function where the white point is the town's health center.

disease was predicted to have a population size of estimated cases and controls with which the correction factor can be determined. As a result, the value of ρ would be 1.346. Using this value in Eq. 2, the fixed prevalence was 0.191. The effect of this correction factor can be seen graphically in the change between Fig. 2 and 3.

A few months after the health campaign, a second data collection was carried out in the town in order to find a greater number of positive cases. In this, it was possible to increase the sampling coverage by 18 percentage points (from 42.5% to 60.5%). As result, hydatidosis prevalence decrease by 0.7 percentage points (from 0.241 to 0.234). It is important to show that the fixed prevalence is not closer to the prevalence after increasing coverage, than to the initial prevalence. The result of the study shows a overestimation in the prevalence, even though increasing the sample size.

Acknowledgments

The authors would like to thank the support of FONDECYT and Cayetano Heredia Peruvian University to make this paper possible

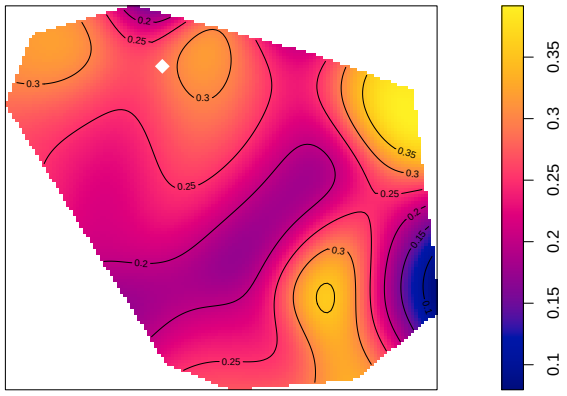

Fig 2. Biased risk. Risk before spatial correction.

References

1. Shaw A, Rushton J, Roth F, Torgerson PR. DALYs, dollars and dogs: how best to analyse the economics of controlling zoonoses. Revue scientifique et technique (International Office of Epizootics). 2017;36(1):147–161.
2. Santivañez SJ, Naquira C, Gavidia CM, Tello L, Hernandez E, Brunetti E, et al. Factores domiciliarios asociados con la presencia de hidatidosis humana en tres comunidades rurales de Junín, Perú. Revista Peruana de Medicina Experimental y Salud Pública. 2010;27:498–505.
3. Moro PL, Budke CM, Schantz PM, Vasquez J, Santivañez SJ, Villavicencio J. Economic Impact of Cystic Echinococcosis in Peru. PLOS Neglected Tropical Diseases. 2011;5:1–6.
4. Celentano DD, Mhs S, Szklo M. Gordis. Epidemiología. Elsevier; 2019.
5. Scheaffer R, Mendenhall W, Ott L, Gerow KG. Elementary Survey Sampling. Cengage Learning; 2011.
6. El-Malki HO, Souadka A, Benkabbou A, Mohsine R, Ifrine L, Abouqal R, et al. Radical versus conservative surgical treatment of liver hydatid cysts. BJS Society. 2014;10(6):669–675.
7. Bailly L, Daurès JP, Dunais B, Pradier C. Bayesian estimation of a cancer population by capture-recapture with individual capture heterogeneity and small sample. BMC medical research methodology. 2015;15:39.
8. Royle JA, Chandler RB, Sollmann R, Gardner B. Spatial Capture-recapture. Academic Press; 2014.

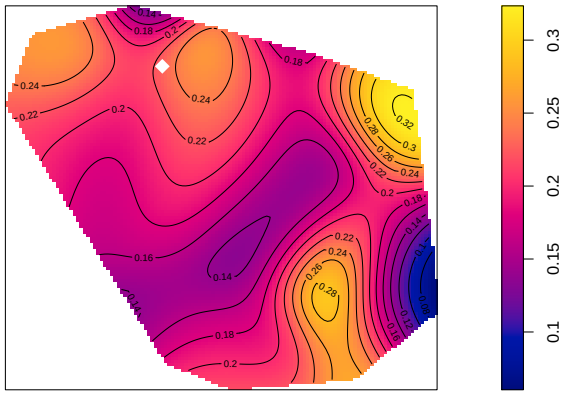

Fig 3. Unbiased risk. Risk after spatial correction.

9. Braeye T, Verheagen J, Mignon A, Flipse W, Pierard D, Huygen K, et al. Capture-recapture estimators in epidemiology with applications to pertussis and pneumococcal invasive disease surveillance. PloS one. 2016;11(8):e0159832.

10. Group WIW, et al. International classification of ultrasound images in cystic echinococcosis for application in clinical and field epidemiological settings. Acta tropica. 2003;85(2):253–261.

11. Davelois K, Escalante H, Jara C. Rendimiento diagnóstico del Western Blot para detectar simultáneamente anticuerpos en pacientes con cisticercosis, hidatidosis y fascioliasis humana. Revista peruana de medicina experimental y salud publica. 2016;33:616–624.

12. Baddeley A, Rubak E, Turner R. Spatial point patterns: methodology and applications with R. CRC press; 2015.

13. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York; 2001.

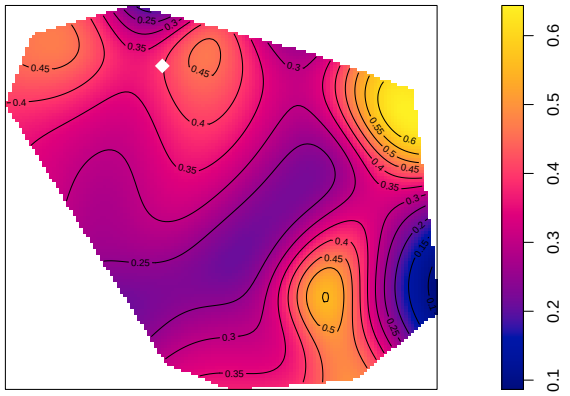

Fig 4. Empirical odds function. Proportion between cases intensity function and controls intensity function where the white point is the town's health center.