USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
Seminario de Iniciación á Investigación
Instituto de Matemáticas

Modelización Onomástica

María José Ginzo Villamayor
Departamento de Estatística e Investigación Operativa

25 de novembro de 2015

Resumo

Os apelidos poden ser utilizados como unha fonte de información para caracterizar a poboación dunha rexión, dado que a análise dos patróns que se observan na distribución dos apelidos reflicte aspectos importantes dos movementos poboacionais. As investigacións desenvolvidas no contexto de estudo, ata a data, non teñen en conta a dimensión espacial e espazo–temporal da evolución dos apelidos; por iso, este traballo céntrase na introdución de métodos estatísticos para o procesamento de datos e o modelado en xeolingüística, especificamente, nos apelidos de Galicia.

Introdución

O obxectivo principal deste traballo é o modelado espacial e espazo-temporal de patróns de apelidos en Galicia. Fixando rexións administrativas, como por exemplo, concellos, pode facerse uso de métodos espaciais e espazo–temporais para a análise de datos de conteo que permitan modelar o patrón subxacente á evolución dos apelidos. Estes métodos serán útiles para caracterizar patróns de evolución dos apelidos formulados mediante modelos xerárquicos. Para o axuste de modelos xerárquicos neste contexto faise uso da metodoloxía INLA (Integrated Nested Laplace Approximation) proposta por [5] que se aplica aos datos dos apelidos en Galicia, proporcionados polo Instituto de Estatística de Galicia (censo de 2011, http://www.ige.eu/).

Neste traballo analízase o patrón espacial, por concellos, dos apelidos "Crujeiras", "Ginzo" e "Rodríguez" en Galicia para o ano 2011. Ademais, faise unha análise espazo–temporal para o apelido "Ginzo". Para este análise usánse os datos do Padrón continuo de poboación dende o ano 2010 ata 2015.

Datos

Os datos dos que se dispón son os do censo de tódolos Concellos de Galicia para o ano 2011. Neste traballo estudaranse tres apelidos de forma espacial: "Crujeiras", "Ginzo" e "Rodríguez", só cando as persoas o levan no primeiro apelido.

Palabras Clave: Besag-York-Mollié; inferencia Bayesiana; INLA; onomástica.

- **Crujeiras**[1]**:** 488 persoas, localízase principalmente na comarca do Barbanza, concretamente no concello de Ribeira. Crujeiras é a castelanización de Cruxeiras, e a maior parte dos seus portadores proceden do lugar Cruxeiras concello de Ribeira.
- **Ginzo**[1]**:** 130 persoas, localízase principalmente na comarca da Mariña Oriental, máis precisamente na Pontenova. Ginzo é a castelanización de Xinzo, topónimo de orixe prerromana (med. Genitio) que se repite en 12 lugares de Galicia, amais de dar nome a un concello (Xinzo de Limia).
- **Rodríguez**[1]**:** 118209 persoas. No ano 2011 era o apelido máis común, repartido por toda Galicia. Rodríguez é o máis frecuente dos apelidos galegos, presente en tódolos concellos de Galicia, con case 240.000 ocorrencias se temos en conta primeira ou segunda posición no apelido. É de tipo patronímico con sufixo "-ez", que era o procedemento unsado na Idade Media para significar "fillo de", porque daquela os apelidos non eran hereditarios, e cambiaban de xeración en xeración.

No período 2000–2015 o número medio de persoas que levan o apelido Ginzo de forma indistinta no primeiro ou no segundo lugar é de 261, cun comportamento bastante uniforme no período. Salientar que nos anos 2000 e 2015 son nos que menos persoas levan o apelido (252) e pola contra no ano 2008 é o ano que máis se rexistran (267).

Metodoloxía

Nas últimas décadas, a dispoñibilidade de datos espaciais e espazo–temporais aumentou de forma considerable, principalmente debido ao avance das ferramentas computacionais que permiten recoller os datos en tempo real.

A análise da distribución xeográfica e temporal dos apelidos permite estudar a variabilidade espacial e temporal da estrutura das poboacións humanas. Deste xeito, os apelidos poden empregarse como fonte de información das características da poboación. Ademais, a análise de patróns de apelidos proporciona información dinámica dos movementos de poboación. En [3] móstrase como a través do estudo dos apelidos mediante medidas de isonimia combinadas con ferramentas de análise clúster se poden obter mapas de apelidos que reflicten o proceso de urbanización das zonas rurais ás urbanas, entre outros. Preséntanse a continuación os dous modelos estudados, espacial e espazo–temporal, empregados neste traballo.

Modelización espacial

Para analizar o patrón espacial dos apelidos en Galicia, considerando o efectos da covariable número de habitantes por concellos, axustouse o modelo proposto en

[1]**Dicionario dos apelidos galegos**, que se está a elaborar na sección de Onomástica do Instituto da Lingua Galega (USC) dirixido por Ana Isabel Boullón Agrelo.

[2] (modelo BYM), adaptado a este contexto. Un dos supostos deste modelo é que o log-risco[2] pódese descompoñer como suma dunha compoñente espacial estruturada e un erro aleatorio, pero tamén se pode incluír o efecto suave dalgunha covariable. En primeiro lugar, para formular o modelo BYM, necesitamos definir cal vai ser a nosa variable de conteo:

Y_i = número de persoas con ese apelido no concello i, para cada $i = 1, \dots, n$.

Este proceso de conteo será modelado a través dun modelo Poisson–LogNormal (ver [1] páx. 162). É dicir, $Y_i|\eta_i \sim \text{Pois}(E_i \exp(\eta_i))$ onde E_i é a poboación en risco (estimación da poboación que leva o apelido en estudo), η_i (os riscos log–relativos) un predictor lineal e as variables Y_i condicionalmente en η_i son independentes. O campo latente η_i terase en conta para modelar a estrutura subxacente e recoller a variabilidade espacial. Unha formulación sinxela vén dada por: $\eta_i = \mu + f_s(s_i) + f_u(s_i)$, onde s_i é o centroide de cada concello e f_s e f_u denotan os efectos espaciais estruturado e o non estruturado, respectivamente. Para f_s, imporase un campo aleatorio Gaussiano de Markov (GMRF) intrínseco ([4]). Para f_u, considerase o proceso de ruído branco que representa a "sobredispersión" que poden presentar os concellos:

- Denotando por $z(s_i) \equiv f_s(s_i), i = 1, \dots, n$, Z é un GMRF, (véxase a Sección 2.2 no libro de [4] para unha definición precisa de GMRF). Os $z(s_i)$, $z(s_j)$ con $i \neq j$ son dependentes con estrutura de Markov e seguen unha distribución $N(0, \tau_2)$.
- Por outra banda, denótase por $w(s_i) \equiv f_u(s_i), i = 1, \dots, n$, W é ruido branco. Os $w(s_i), i = 1, \dots, n$ son independentes con distribución $N(0, \tau_1)$.

Modelización espazo–temporal

Investigando só o patrón espacial dos apelidos non nos permite concluír nada sobre outra das compoñentes de variación, a temporal, que poder ser igualmente de interese. O modelo anterior pode estenderse facilmente ao caso espazo–temporal incluíndo o tempo. A variable resposta para un concello i será:

Y_{it} = o número de persoas con ese apelido no concello i no tempo t,

que será observada nos n concellos e en T instantes do tempo. O modelo espacial anterior esténdese ao permitir a compoñente temporal quedando: $Y_{it}|\eta_i \sim \text{Pois}(E_{it} \exp(\eta_{it}))$. A formulación que seguirá neste caso o campo latente é: $\eta_{it} = \mu + f_s(s_i) + f_u(s_i) + f_T(t)$, con $t = 1, \dots, T$, onde en $f_T(t)$ se especifica a estrutura temporal. Dita estrutura pode corresponder a unha compoñente lineal no tempo, a unha compoñente suave ou ben a un proceso que inclúa correlación temporal.

[2] Estes modelos empréganse no ámbito da epidemioloxía. Aquí log–risco enténdese como a poboación susceptible de posuír o apelido en estudo, en palabras epidemiolóxicas, o risco de "padecer" un determinado apelido.

Inferencia Bayesiana con INLA

O axuste dos modelos realizouse empregando o método baseado na aproximación por integradores de Laplace aniñados INLA[3], proposto en [4].

INLA propociona unha ferramenta rápida e útil para axustar modelos gaussianos latentes (os procesos que rixen f_s, f_u e f_T teñen distribucións Gaussianas), incluíndo modelos con estrutura temporal ou espacial nun contexto Bayesiano. Neste traballo farase uso deste metodoloxía para axustar un modelo aos datos de tres apelidos en Galicia, Crujeiras, Ginzo e Rodríguez e detallarase como se fixo o axuste con `R-INLA`[4] (dispoñible en www.r-inla.org).

Os obxectivos da inferencia Bayesiana son as distribucións marxinais a posteriori para cada compoñente do vector de parámetros. O procedemento INLA calcula a aproximación numérica das distribucións a posteriori que sexan de interese, baseados no método de aproximación de Laplace.

Para levar a cabo a inferencia Bayesiana hai que especificar os hiperparámetros do efecto espacial a priori e da parte aleatoria. Os hiperparámetros son a precisión do τ_1 no modelo `iid`[5] e a precisión τ_2 do modelo `besag`[6], como $\theta = (\log \tau_1, \log \tau_2)$. Sobre estes parámetros considéranse unhas distribucións a priori, log–gamma neste caso, con parámetros iniciais (1,0.00005). A análise de sensibilidade do modelo á elección do parámetro das distribucións a priori pode facerse en base ao criterio Criterio de Información de Desviación (DIC). Este criterio baséase na desviación Bayesiana. É unha xeneralización do Criterio de Información de Akaike (AIC) e do Criterio de Información Bayesiano (BIC). O criterio DIC emprégase, especialmente, sobre distribucións a posteriori obtidas mediante métodos Markov chain Monte Carlo (MCMC).

Ademais é un criterio que proporciona bos resultados para modelos xerárquicos como os aplicados neste traballo. Os modelos que teñen un valor DIC máis pequeno son os preferidos. O DIC penaliza tanto polo axuste como pola complexidade do modelo. O axuste soe ser mellor ao introducir un número maior de parámetros ao modelo, pero isto compénsase no criterio DIC cunha penalización segundo o número de parámetros buscando un equilibrio entre a bondade do axuste e a complexidade do modelo. Para a súa construción débese definir a desviación: $D(\lambda) = -2 \ln (p(Y|\lambda))$, onde Y representa a variable resposta, λ os parámetros específicos de cada modelo (os parámetros que aparecen no BYM son μ, τ_1 e τ_2). Así mesmo pode obterse a desviación media a posteriori como: $\bar{D} = E[D(\lambda)]$. Por outra banda a desviación Bayesiana das medias a posteriori é $\hat{D} = D[E(\lambda)]$. Deste xeito temos que o criterio DIC ven definido por: $DIC = 2\bar{D} - \hat{D}$.

[3] Bayesian computing with INLA!, (2009) URL: http://www.r-inla.org/.

[4] `R-INLA` é un paquete de `R` que implementa a aproximación da inferencia Bayesiana usando INLA (Martino e Rue, 2010a).

[5] ver http://www.math.ntnu.no/inla/r-inla.org/doc/latent/indep.pdf

[6] ver http://www.math.ntnu.no/inla/r-inla.org/doc/latent/besag.pdf

Resultados e conclusións

Resultados da modelización espacial con INLA

Para o ano 2011 en Galicia hai 488 persoas que levan o apelido Crujeiras, 130 o apelido Ginzo e 118209 o apelido Rodríguez. Fíxose un axuste de acordo ao exposto na Sección referida a Modelización espacial tomando os hiperparámetros a priori que emprega por defecto en `inla`, xa que o emprego de outros non melloran os resultados. O grafo de Galicia non é un conxunto conexo, senón que está formado por dúas compoñentes conexas: a Illa de Arousa e os 314 restantes concellos integrados na Península Ibérica. Isto hai que indicarllo ao modelo. A función `inla` devolve un obxecto da clase `inla`, que contén unha serie de elementos que poden ser explorados, como son a media, desviación típica e cuantiles. Na Figura 1 móstrase o patrón espacial dos apelidos Crujeiras, Ginzo e Rodríguez tendo en conta o número de habitantes por concello, e representa a media a posteriori do efecto estruturado. Por outra banda, na Figura 2 móstrase a parte aleatoria non estruturada.

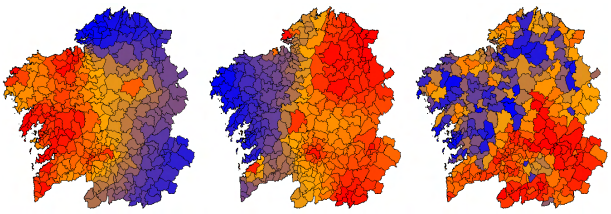

Figura 1: Media a posteriori da parte estruturada do campo latente η_i. De esquerda a dereita: para os apelidos Crujeiras, Ginzo e Rodríguez dos datos do censo 2011.

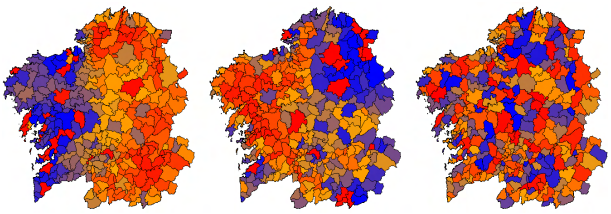

Figura 2: Media a posteriori do efecto non estruturado do campo latente η_i. De esquerda a dereita: para os apelidos Crujeiras, Ginzo e Rodríguez dos datos do censo 2011.

Resultados da modelización espazo–temporal con INLA

Neste apartado fíxose un axuste sinxelo de acordo ao exposto na Sección referida a Modelización espazo–temporal. Eliximos como parámetro para o efecto temporal o `prior.iid=c(1,1)` e axustamos o modelo ao segundo conxunto de datos. Neste caso emprégase como "poboación en risco" o 20 % da poboación total do concello, facendo

unha estimación naqueles concellos que o apelido Ginzo está máis presente. Tanto o `intercepto` como o coeficiente asociado á variable ano, que recolle a compoñente temporal, son significativos.

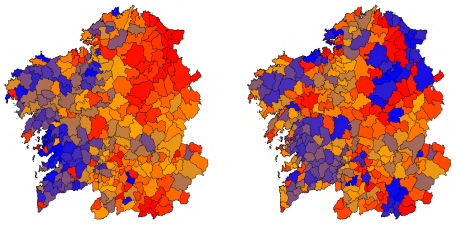

Figura 3: Esquerda: Media a posteriori da parte non estruturada do campo latente η_i para o axuste do modelo espazo–temporal. Dereita: Efecto debido ó tempo.

A conclusión que se extrae deste estudo unha vez aplicado o modelo espazo–temporal fronte ao modelo espacial é que os mapas das distribucións dos apelidos son similares. Quizás poida dicirse que nas veciñanzas desas 261 persoas que levan o apelido Ginzo non hai moito movemento nin no espazo nin ao longo do tempo. Seguramente se tivéramos máis anos poderíamos obter resultados máis interesantes, xa que tendo en conta a idade das persoas e facendo cortes temporais como en [3] mediante técnicas de análise clúster pódese detectar o proceso de urbanización en Galicia ou mesmo o mapa das dioceses de Galicia (antes do proceso urbanización o mapa dos apelidos de Galicia distribuíase da mesma forma que o mapa das dioceses).

Bibliografía

[1] Banerjee S., Carlin, B. E., Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability

[2] Besag, J., York, J. e Mollié, A. (1991). *Bayesian image restoration with two applications in spatial statistics*. Annals of the Institute of Statistical Mathematics, **43**, pp. 1–59.

[3] Ginzo-Villamayor, M.J.; Crujeiras, R. M. e Sousa Fernández, X. (2013). *Surname patterns in Galicia*. Libro de Actas do congreso. XI Congreso Galego de Estatística e Investigación de Operacións. A Coruña (España).

[4] Rue, H., e Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman & Hall, CRC Press.

[5] Rue, H., Martino, S. e Chopin, N (2009). *Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations*. Journal of the Royal Statistical Society, Series B, **71**, pp. 319–392.