

CALIBRACIÓN POR SIMULACIÓN DEL CONTRASTE DE ORDEN MARKOVIANO EN TABLAS DE CONTINGENCIA RALAS: ANÁLISIS EMPÍRICO SOBRE UN MILLÓN DE TRASBORDOS DEL TRANSPORTE PÚBLICO

José Ignacio Colomé Vita

Departamento de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina

ignacio.colome@unsj-cuim.edu.ar

El sistema SUBE de San Juan registra el ascenso de cada pasajero: de abril de 2023 se reconstruyen 1.015.178 tripletas de trasbordos consecutivos entre líneas de transporte público, sobre hasta 116 estados. Modelarlas como una cadena de Markov exige decidir antes si la secuencia es de primer orden, es decir, si cada transición depende solo del estado actual; el procedimiento clásico es el contraste de razón de verosimilitudes de Anderson y Goodman, frente a la alternativa de segundo orden. Bajo la hipótesis nula, $\Lambda = -2 \log \lambda$ sigue asintóticamente una ji cuadrado con grados de libertad contados sobre las dimensiones observadas, de modo que Λ/gl debería promediar la unidad. Con ese número de estados la tabla de segundo orden tiene más casilleros que observaciones: hasta un 98,8% de sus celdas tiene frecuencia esperada menor que 5, muy por encima del 20 % que tolera la regla habitual. La aproximación asintótica se degrada en ese régimen, hecho conocido pero no cuantificado para este contraste y característico de los registros transaccionales masivos. Aquí se mide por simulación ese apartamiento. Con m estados y N tripletas, se denomina línea base b al valor medio de Λ/gl sobre cadenas generadas como de primer orden por construcción, y c al número esperado de observaciones por casillero:

$$b = E_{H_0}[\Lambda/\text{gl}], \quad c = N/m^3.$$

La teoría sitúa b en la unidad. Sobre una grilla con m entre 3 y 114 estados y N entre 10.000 y 1.000.000, la línea base promedia 0,99 mientras las celdas ralas no superan el 85 %; baja a 0,81 entre el 85 % y el 95 % y a 0,48 por encima, con un mínimo de 0,31. Configuraciones muy distintas de m y N producen líneas base semejantes al ordenarlas por c : la ralitud predice el sesgo mucho mejor que el tamaño muestral. Con la línea base por debajo de la unidad el contraste se vuelve conservador y pierde potencia, de modo que deja de rechazar violaciones que la línea base sí registra. Sobre los ocho subconjuntos de trasbordos examinados, definidos por franja horaria, jornada, tipo de día y tipo de trasbordo, la línea base de cada uno, medida y no interpolada, se ubica entre 0,53 y 0,72. Un control sobre una subcadena densa de tres líneas arroja 0,86 y no la unidad; el apartamiento se explica por sus doce grados de libertad y es casi cuatro veces menor que el del régimen ralo. Calibrar contra la línea base simulada, y no contra la unidad teórica, invierte el diagnóstico en tres de los ocho subconjuntos: bajo el criterio asintótico parecían de primer orden y contra su propia línea base revelan un exceso de entre el 23 % y el 39 %. Se obtiene así un procedimiento de calibración reproducible, previo al contraste sobre tablas de contingencia ralas. Este trabajo se enmarca en las actividades de investigación del proyecto PIC E1212.

Trabajo en conjunto con Leonel Ganga (Departamento de Matemática y Química, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina).

Referencias

- [1] T. W. Anderson y L. A. Goodman, Statistical inference about Markov chains, The Annals of Mathematical Statistics, vol. 28, n.º 1, pp. 89-110, mar. 1957. doi: 10.1214/aoms/1177707039.
- [2] K. J. Koehler, Goodness-of-fit tests for log-linear models in sparse contingency tables, Journal of the American Statistical Association, vol. 81, n.º 394, pp. 483-493, 1986. doi: 10.1080/01621459.1986.10478294.
- [3] S. J. Haberman, A warning on the use of chi-squared statistics with frequency tables with small expected cell counts, Journal of the American Statistical Association, vol. 83, n.º 402, pp. 555-560, 1988. doi: 10.1080/01621459.1988.10478632.

- [4] J. Besag y P. Clifford, Generalized Monte Carlo significance tests, *Biometrika*, vol. 76, n.º 4, pp. 633-642, 1989. doi: 10.1093/biomet/76.4.633.
- [5] P. Billingsley, Statistical methods in Markov chains, *The Annals of Mathematical Statistics*, vol. 32, n.º 1, pp. 12-40, 1961. doi: 10.1214/aoms/1177705136.
- [6] X.-L. Ma, Y.-H. Wang, F. Chen y J.-F. Liu, Transit smart card data mining for passenger origin information extraction, *Journal of Zhejiang University SCIENCE C (Computers Electronics)*, vol. 13, n.º 10, pp. 750-760, oct. 2012. doi: 10.1631/jzus.C12a0049.
- [7] Y. Zhou, L. Wang, R. Zhong y Y. Tan, A Markov chain based demand prediction model for stations in bike sharing systems, *Mathematical Problems in Engineering*, vol. 2018, art. 8028714, en. 2018. doi: 10.1155/2018/8028714.
- [8] R. Besenczi, N. Bátfai, P. Jeszenszky, R. Major, F. Monori y M. Ispány, Large-scale simulation of traffic flow using Markov model, *PLOS ONE*, vol. 16, n.º 2, art. e0246062, feb. 2021. doi: 10.1371/journal.pone.0246062.