

Artículo de investigación

## Evaluación de la agregación en los modelos de Edad-Periodo-Cohorte (APC) con un enfoque Bayesiano

### Evaluation of aggregation in Age-Period-Cohort (APC) models with a Bayesian approach

Simón Cuartas Rendón<sup>1</sup>, Isabel Cristina Ramírez Guevara<sup>1</sup>✉ y Johnatan Cardona Jiménez<sup>1</sup>

<sup>1</sup>Universidad Nacional de Colombia Sede Medellín.

**Recepción:** 15-enero-2025    **Aceptado:** 20-mayo-2025    **Publicado:** 20-julio-2025

**Cómo citar:** Ramírez Guevara, I. C., Cardona Jiménez, J., & Cuartas Rendón, S. . (2025). Evaluación de la agregación en los modelos de Edad-Periodo-Cohorte (APC) con un enfoque Bayesiano. *Ciencia En Desarrollo*, 16(2). doi: 10.19053/uptc.01217488.v16.n2.2025.17094

#### Resumen

Los modelos de regresión Poisson de Edad-Periodo-Cohorte, APC por su sigla en inglés, son utilizados en estudios de epidemiología para estimar el efecto de cada uno de estos factores en las tendencias de las razones de incidencia y mortalidad de diferentes enfermedades. Generalmente, se tiene la información sobre los casos observados o la mortalidad de una enfermedad resumidos en una tabla con dos entradas: el grupo de edad y el período calendario en que se registró el evento de interés. Uno de los retos que un investigador debe enfrentar en los diferentes contextos y escenarios que requieren aplicaciones de modelos APC es la agregación de los datos. En la mayoría de los estudios, la información disponible sólo se encuentra agregada ya sea por grupos de edades, por periodos o por ambas características. En este sentido sería pertinente evaluar el efecto que tiene esta agregación de la información tanto en la estimación de los efectos del modelo como en las predicciones obtenidas con éste. Se podría inferir que esta agregación, usualmente por periodos de 5 años o quinquenios, debe afectar de manera significativa la calidad de las predicciones, pues en este proceso se pueden perder tendencias o patrones estacionales. Así, en este trabajo se realiza un análisis sobre el impacto que tiene la agregación de datos de las incidencias (por periodos) en las proyecciones o predicciones futuras. En particular, se hace un estudio comparativo entre las tasas de mortalidad por tuberculosis para el caso de Colombia y se muestran proyecciones para tres agregaciones del periodo: quinquenios, trienios y anualidades.

**Palabras Clave:** Modelos de Edad-Periodo-Cohorte (APC), Enfoque Bayesiano, Análisis de agregación

#### Abstract

Poisson Regression models of Age-Period-Cohort (APC) are employed in epidemiological studies to estimate the impact of each of these factors on the trends in incidence and mortality rates for various diseases. Typically, information on observed cases or mortality from a disease is summarized in a table with two entries: age group and calendar period in which the event of interest was recorded. One of the challenges researchers face in various contexts and scenarios requiring APC model applications is data aggregation. In most studies, the available information is aggregated either by age groups, periods, or both characteristics. In this regard, it would be relevant to assess the impact of this data aggregation on both the estimation of model effects and the predictions obtained from it. It could be inferred that this aggregation, usually done in 5-year periods or quinquennia, may significantly affect the quality of predictions, as trends or seasonal patterns may be lost in this process. Thus, this work conducts an analysis on the impact of aggregating incidence data (by periods) on future projections or predictions. In particular, a comparative study is carried out on tuberculosis mortality rates for the case of Colombia, and projections are shown for three period aggregations: five-year periods, three-year periods, and annuities.

**Keywords:** Age-period-cohort (APC) models, Bayesian approach, Aggregation analysis

## 1. Introducción

Los modelos APC suelen utilizarse en estudios de epidemiología para analizar las tendencias de la incidencia y la mortalidad de diferentes enfermedades y, en general, pueden aplicarse a cualquier situación en la que se pretenda valorar el efecto temporal de la ocurrencia de un evento. Una de las motivaciones principales de analizar los modelos APC es estimar el efecto de cada uno de estos factores por separado en la evolución de las tasas. Generalmente, se tiene la información sobre los casos observados o la mortalidad de una enfermedad resumidos en una tabla con dos entradas: el grupo de edad y el período calendario en que se registró el evento de interés. El efecto de la edad representa el cambio en las tasas asociado a la edad. El período y la cohorte son efectos que explican cambios en las razones asociados con el tiempo. El efecto período representa cambios en las razones debidos a factores localizados en un momento del tiempo, que influyen en todos los grupos de edad simultáneamente y están relacionados con la tecnología. El efecto cohorte se asocia a factores que afectan a una generación y provocan cambios diferentes en las tasas de magnitud en sucesivos grupos de edad y períodos, y que se relacionan con estándares culturales y de comportamiento. Un ejemplo del efecto del período es un cambio en un tratamiento, que podría reducir la mortalidad para todos los grupos de edad al mismo tiempo. Los efectos de la cohorte están asociados a hábitos de larga duración [1].

La relación lineal exacta existente entre estos tres efectos de estudio hace que los parámetros del modelo completo no puedan estimarse, lo que se denomina no identificabilidad. Múltiples soluciones para este problema se han sugerido; la mayoría de ellas proponen restricciones sobre los parámetros. [2] muestran que esto se puede resolver haciendo una parametrización del modelo en la cual los parámetros tienen una interpretación en términos del riesgo relativo, tomando un período o una cohorte como referencia. Otro enfoque para encontrar un único grupo de parámetros es igualar dos de los efectos para uno de los factores del modelo. [3] sugieren una variación de este enfoque que supone que la media de diferencias sucesivas sea cero.

En la mayoría de los estudios, la información disponible sólo se encuentra agregada ya sea por grupos de edades, por períodos o por ambas características. En este sentido sería pertinente evaluar el efecto que tiene esta agregación de la información tanto en la estimación de los efectos del modelo como en las predicciones obtenidas con éste. Se podría inferir que esta agregación, usualmente por períodos de 5 años o quinquenios, debe afectar de manera significativa la calidad de las predicciones, pues en este proceso se pueden perder tendencias o patrones estacionales. Así, en este trabajo se realiza un análisis sobre el impacto que tiene la agregación de datos de las incidencias (por períodos) en las proyecciones o predicciones futuras. En particular, se hace un estudio comparativo de tres agregaciones de período distintas: anualidades, trienios y quinquenios, para estudiar el efecto de estas agregaciones en las predicciones de las tasas de mortalidad por tuberculosis para el caso de Colombia.

Este artículo se encuentra organizado de la siguiente manera:

en la sección 2 se describe el modelo APC, sus supuestos y algunas soluciones al problema de identificabilidad del modelo. En la sección 3 se presenta la aplicación. En la sección 4 se muestran los resultados. Finalmente, la sección 5 se dan las conclusiones.

## 2. Modelo APC

La información que se considera en estos modelos son los casos observados o mortalidad de una enfermedad resumidos en una tabla de dos entradas: el grupo de edad y el período en que se registró el evento. Típicamente, el ancho del intervalo para edad y período son iguales, por lo tanto si los períodos son divididos en intervalos de cinco años, la edad también. Los grupos de edad se representan por  $a (= 1, \dots, A)$  y los períodos por  $p (= 1, \dots, P)$ . La cohorte es representado por  $c (= 1, \dots, C)$ . Existe una relación entre estos tres factores dada por  $c = p - a + A$ . La forma general del modelo de regresión que incluye los tres factores está dada por una estructura aditiva que incluye los efectos de edad, período y cohorte.

Siguiendo la estructura definida en [4], los supuestos del modelo son los siguientes:

- El número de casos en el grupo de edad  $a$  en el período  $p$  y la cohorte  $c$  se denota por  $y_{apc}$ , la cual se supone es una realización de una variable aleatoria Poisson, con media  $\lambda_{apc}$ . Se tiene que  $a = 1, \dots, A$ ,  $p = 1, \dots, P$  and  $c = 1, \dots, A + P - 1$ . El número de grupos de edad es  $A$  y el número de períodos es  $P$ .
- El número de personas en riesgo en el grupo de edad  $a$ , en el período  $p$  y cohorte  $c$  ( $n_{apc}$ ) es un valor fijo conocido.
- El logaritmo de la razón esperada entre  $\lambda_{apc}$  y  $n_{apc}$  es una función lineal de los efectos del grupo de edad, el período y la cohorte de nacimiento:

$$\log\left(\frac{\lambda_{apc}}{n_{apc}}\right) = \mu + \alpha_a + \beta_p + \gamma_c \quad (1)$$

donde los efectos de la edad se representan por  $\alpha_a$ , los del período por  $\beta_p$ , y los efectos de la cohorte por  $\gamma_c$ .

Si se conocen dos de los valores de  $a$ ,  $p$ , o  $c$ , el tercero puede ser determinado dado que  $c = p - a + A$ . Esta correspondencia entre estos factores es llamada dependencia lineal. En análisis de regresión, cuando las variables regresoras son linealmente dependientes, no es posible atribuir efectos separados a cada uno de estos factores, esto es llamado como problema de no identificabilidad. Así que se deben aplicar restricciones adicionales a los parámetros para obtener una solución. Un enfoque consiste en establecer  $\alpha_1 = \beta_1 = \gamma_1 = 0$ . Otra alternativa es hacer que la suma de los efectos sea igual a cero,  $\sum_a \alpha_a = 0$ ,  $\sum_p \beta_p = 0$  y  $\sum_c \gamma_c = 0$ . Desafortunadamente, estas restricciones no resuelven completamente el problema (ver [5]). El problema de no identificabilidad ha sido ampliamente estudiado, ver por ejemplo [8, 9, 10]

En [6] predicen razones de mortalidad de cáncer de pulmón ajustando un modelo APC con un enfoque Bayesiano. Especifican un modelo a priori autoregresivo gaussiano hacia adelante para suavizar los efectos en cada escala de tiempo y extrapolar los efectos del periodo y la cohorte. De acuerdo con este modelo, cada punto, excepto los dos primeros en cada escala de tiempo, se predice por una extrapolación lineal de sus dos predecesores inmediatos, más una error aleatorio de una distribución normal con media cero. Las precisiones de estas distribuciones normales se pueden representar como unos hiperparámetros para los cuales la distribución a priori refleja la creencia previa del suavizamiento de los parámetros.

En [7] se implementa un modelo APC Bayesiano a múltiples bases de datos. Define como distribución a priori para los parámetros de precisión de los efectos la Gamma (0.001,0.001). Realiza proyecciones suponiendo que los dos últimos periodos no fueron observados y compara el desempeño del modelo Bayesiano con modelos bajo enfoques diferentes.

Las razones ajustadas y proyectadas se obtienen sumando las efectos suavizados de edad, periodo y cohorte de acuerdo con:

$$\text{rate}_i = \exp(\alpha_{a_i} + \beta_{p_i} + \gamma_{c_i}) \quad 1 \leq i \leq AP \quad (2)$$

### 3. Aplicación datos de mortalidad por tuberculosis en Colombia

Se realiza el ajuste del modelo APC con verosimilitud Poisson a la base de datos de mortalidad por tuberculosis en Colombia, la cual se obtuvo de las estadísticas vitales que publica el Departamento Administrativo Nacional de Estadística (DANE, <https://www.dane.gov.co>). Se consideró que la población bajo riesgo está integrada por todos los habitantes del territorio nacional de Colombia, estos datos también se obtuvieron del DANE. Se tiene la mortalidad desde 1979 hasta 2019 para dieciséis grupos de edad que van desde el grupo [5 años - 9 años] hasta [80 años y más]. Se hace el ajuste del modelo dejando un periodo por fuera para realizar la evaluación del pronóstico. En particular, cuando la agregación fue por quinquenios, para el ajuste se utilizaron siete quinquenios siendo el primero [1979-1983] y el último [2009-2013] y se pronosticó [2014-2018]. Para los trienios se usaron doce trienios, dejando el trienio [2015-2017] para validar la calidad del pronóstico. Por último, para la agregación anual, se realizó el ajuste incluyendo los años desde 1979 hasta 2017, y se validó el pronóstico del 2018.

Posteriormente se realizan las comparaciones de las razones de mortalidad estimadas con las razones reales para el periodo que se excluyó en el ajuste. Se utilizan como medidas de desempeño de las predicciones el error cuadrático medio (ECM) y el error absoluto medio (MAE), los cuales están dados por:

$$\text{ECM} = \frac{\sum_{a=1}^A (\hat{\lambda}_a - \lambda_a)^2}{A},$$

$$\text{MAE} = \frac{\sum_{a=1}^A |\hat{\lambda}_a - \lambda_a|}{A},$$

donde  $\hat{\lambda}_a$  son las predicciones de las razones de mortalidad de la  $a$ -ésima banda de edad en el periodo de validación. Debido a que se utiliza un enfoque Bayesiano, para comparar el ajuste de los modelos APC de acuerdo con la agregación, se utiliza el criterio de información de devianza (DIC, por sus siglas en inglés) propuesto por [11]. El DIC es el criterio de comparación de modelos más utilizado debido a su facilidad computacional y a que se encuentra implementado en los software donde se ajustan modelos con un enfoque Bayesiano. Con este criterio el mejor modelo es el que presenta un menor DIC.

Siguiendo el enfoque definido en [6] se especifica un modelo a priori autorregresivo Gaussiano para suavizar los efectos para cada escala de tiempo y evitar el problema de no identificabilidad. Con este modelo cada punto (excepto los dos primeros en cada escala) se predice a través de una extrapolación lineal de sus dos predecesores inmediatos, más un error aleatorio de una distribución Normal con media cero. A los dos primeros parámetros de cada efecto se les asigna una distribución a priori no informativa. Por ejemplo, para los  $A$  efectos de edad la distribución a priori es:

$$\begin{aligned} \alpha_1 &\sim \text{Normal}(0, \sigma_\alpha^2); \\ \alpha_2 | \alpha_1 &\sim \text{Normal}(0, \sigma_\alpha^2); \\ \alpha_a | \alpha_1, \alpha_2, \dots, \alpha_{a-1} &\sim \text{Normal}(2\alpha_{a-1} - \alpha_{a-2}, \sigma_\alpha^2), \end{aligned} \quad (3)$$

donde  $\sigma_\alpha^2$  es la varianza del efecto edad. Se especifica como distribución a priori para las varianzas de los efectos,  $\sigma_\alpha^2, \sigma_\beta^2$  y  $\sigma_\gamma^2$ , Gamma-inversa(0.001,0.001).

Se ajusta el modelo corriendo 400,000 iteraciones MCMC con un quemado de 40,000 y se realizan las estimaciones después de haber verificado convergencia de las cadenas. Se utilizó el software JAGS (Just Another Gibbs Sampler) usando los paquetes de R, rjags y R2jags.

### 4. Resultados

En la Tabla 1 se observan las medidas de desempeño del error de la predicción para el periodo de validación de acuerdo a la agregación. El modelo basado en trienios es el que realiza mejores pronósticos, seguido del modelo basado en anualidades. Sin embargo, el modelo que presenta un mejor ajuste es en que el periodo esta agregado por quinquenios, ver Tabla 2, pues es el que presenta un menor DIC. También se observa que conforme mayor es la agregación en los periodos, mejor es la calidad de ajuste del modelo.

Tabla 1: Medidas de desempeño para la predicción

|     | Quinquenios | Trienios | Anualidades |
|-----|-------------|----------|-------------|
| ECM | 2.838       | 1.032    | 2.417       |
| MAE | 1.032       | 0.579    | 0.987       |

Tabla 2: Valores DIC

| Quinquenios | Trienios | Anualidades |
|-------------|----------|-------------|
| 1310.204    | 2289.85  | 5065.258    |

La Figura 1 muestra la estimación del efecto del periodo cuando la agregación es en trienios. Se observa que en los primeros cinco trienios el efecto del periodo es muy similar, sin embargo, a partir del sexto trienio, a medida que aumenta los trienios el efecto del periodo también aumenta. Adicionalmente, esta Figura muestra la predicción del efecto del periodo para el trienio 13 el cual comprende los años [2015-2017], este pronóstico sigue la tendencia positiva del efecto del periodo. De otro lado, cuando la agregación es en quinquenios, no se observa una tendencia en la estimación del efecto del periodo, ver Figura 2, y el intervalo de predicción del efecto para el quinquenio [2014-2018] es más amplio comparado con el intervalo que se obtuvo cuando la agregación fue con trienios.

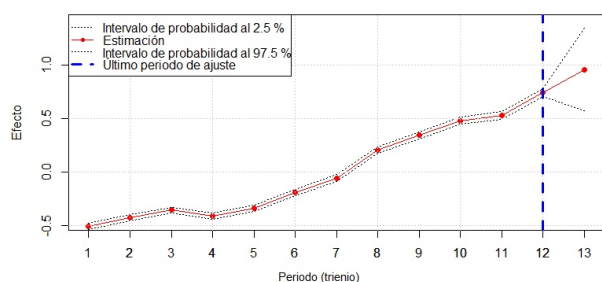


Figura 1: Estimación y predicción del efecto del periodo con intervalos para la agregación en trienios.

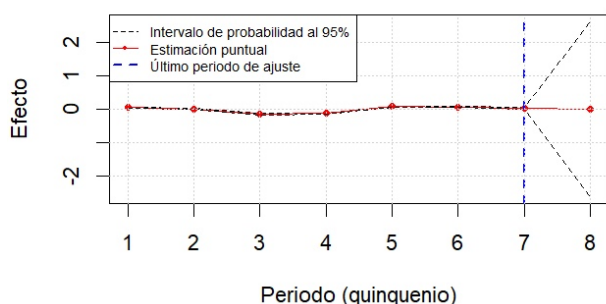


Figura 2: Estimación y predicción del efecto del periodo con intervalos para la agregación en quinquenios.

La Figura 3 ilustra las muertes ajustadas y observadas para la agregación en trienios. Las muertes ajustadas conservan la misma tendencia de las muertes observadas, a medida que aumenta el periodo las muertes por tuberculosis en Colombia disminuyen levemente.

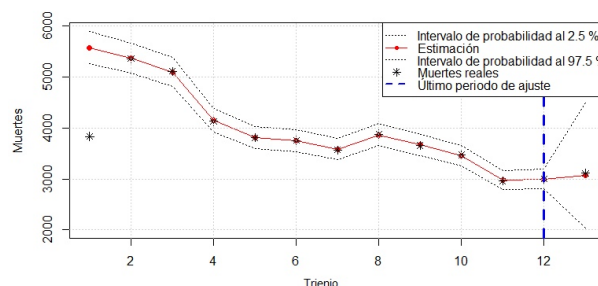


Figura 3: Estimaciones y predicción de la mortalidad por tuberculosis en Colombia con intervalos para el efecto del periodo con agregación en trienios.

## 5. Conclusiones

Al realizar una aplicación del modelo APC con diferentes agregaciones del periodo se observa que esto tiene un efecto tanto en el ajuste como en la predicción. Para el caso en particular de estudio en este artículo, se tiene que las mejores predicciones se obtienen cuando la agregación fue por trienios. En el caso del ajuste se observa que, de acuerdo con el DIC, a medida que aumenta la agregación de periodos, mejor es la calidad del ajuste.

Para obtener conclusiones más generales sobre el efecto de la agregación en el modelo APC, se deja como trabajo futuro un estudio de simulación que involucre diversos comportamientos poblacionales de los datos y así poder concluir cuál es la mejor agregación de los datos para lograr un mejor modelo en términos de ajuste y predicción.

## Referencias

- [1] J. R. González, F. J. Llorca y V. Moreno, "Algunos aspectos metodológicos sobre los modelos edad-periodo-cohort. Aplicación a las tendencias de mortalidad por cáncer", *Gaceta Sanitaria*, vol. 16, no. 3, pp. 267-273, 2002.
- [2] D. Clayton, y E. Schifflers, "Models for temporal variation in cancer rates. I: Age-period and age-cohort models", *Stat Med.*, vol 6, no. 4, pp. 449-67, 1987.
- [3] D. Clayton, y E. Schifflers, "Models for temporal variation in cancer rates. II: Age-period-cohort models", *Stat Med.*, vol 6, no. 4, pp. 469-81, 1987.
- [4] T. R. Holford, "The Estimation of Age, Period and Cohort Effects for Vital Rates", *Biometrics*, vol 39, no. 2, pp. 311-24, 1983.
- [5] T. R. Holford, "Understanding the effects of age, period, and cohort on incidence and mortality rates", *Annu Rev Publ Health*, vol 12, pp. 425-57, 1991.
- [6] C. Berzuini y D. Clayton, "Bayesian analysis of survival on multiple time scales", *Statistics in Medicine*, vol 13, no. 8, pp. 823-38, 1994.
- [7] I. Bray, "Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality

- ", *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol 51, no. 2, pp. 151-64, 2002.
- [8] W. Fu, "Ridge estimator in singular design with application to Age-Period-Cohort analysis of disease rates ", *Communications in Statistics - Theory and Methods*, vol 29, pp. 263-78, 2000.
- [9] W. Fu, "Constrained Estimators and Consistency of a Regression Model on a Lexis Diagram ", *Journal of the American Statistical Association*, vol 111, no. 513 pp. 180-99, 2016.
- [10] E. Fosse y C. Winship, "Bounding Analyses of Age-Period-Cohort Effects ", *Demography*, vol 56, no. 5 pp. 1975-2004, 2019.
- [11] D.J. Spiegelhalter, N. G. Best, B. P. Carlin y A. Van Der Linde, "Bayesian measures of model complexity and fit ", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol 64, no. 4 pp. 583-639, 2002.