



*Casos de Estudio Aplicados al Sector
Seguros y Financiero en Colombia*

Maestría de Analítica y Gerencia de Datos





Pricing

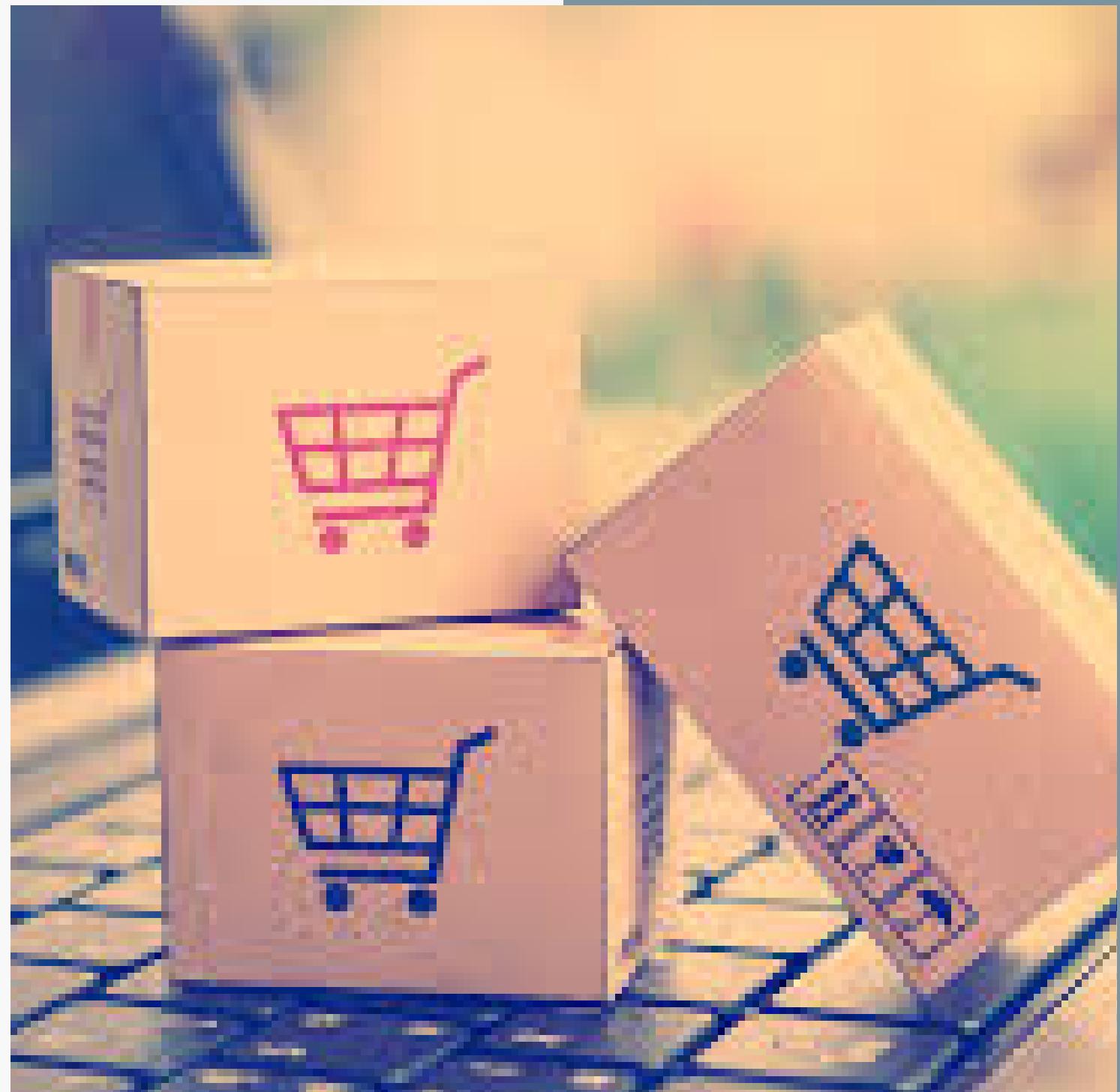
El Pricing es el área dentro de la actuaria mediante la cual se establecen los precios de un producto o servicio. No es más que cuantificar el **mejor precio posible** de este tipo de elementos



Pricing conveniconal

En sectores como el de manufactura, industria, con productos o servicios tangibles típicamente el precio se establece teniendo en cuenta dos aspectos

1. Precio de materiales + precio de transformación + Rentabilidad exigida al producto.
2. ¿Es el precio competitivo con el mercado?, análisis de competidores, distribución del mercado, etc.



Seguros



El precio se basa en un producto con condiciones claras, pero que no se materializa siempre, hay **incertidumbre en el costo del producto.**

¿Sabes cuantos siniestros te va a dar un asegurado?

¿Sabes que tan graves serán estos siniestros?

Seguros



El precio se basa en un producto con condiciones claras, pero que no se materializa siempre, hay **incertidumbre en el costo del producto.**

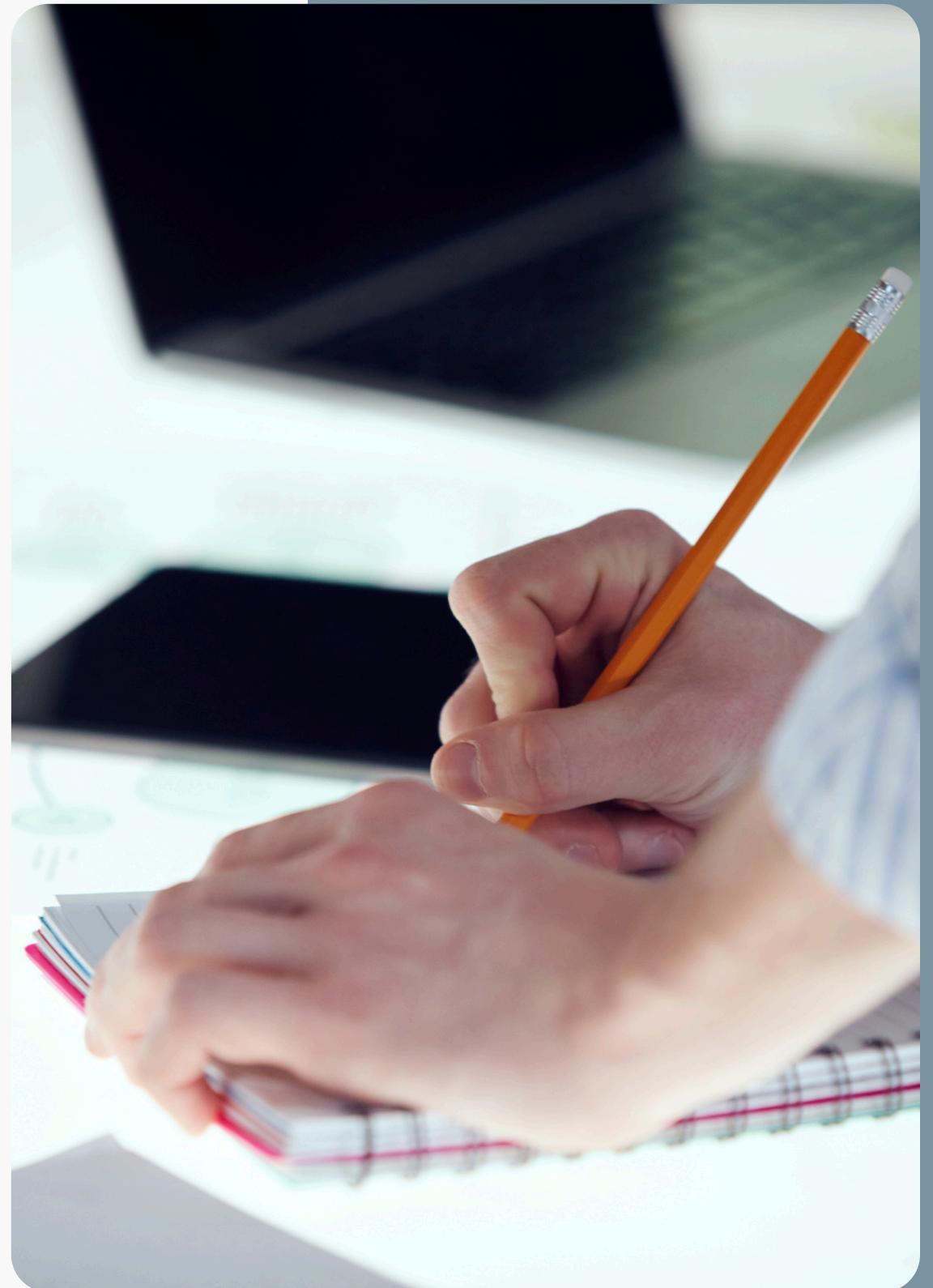
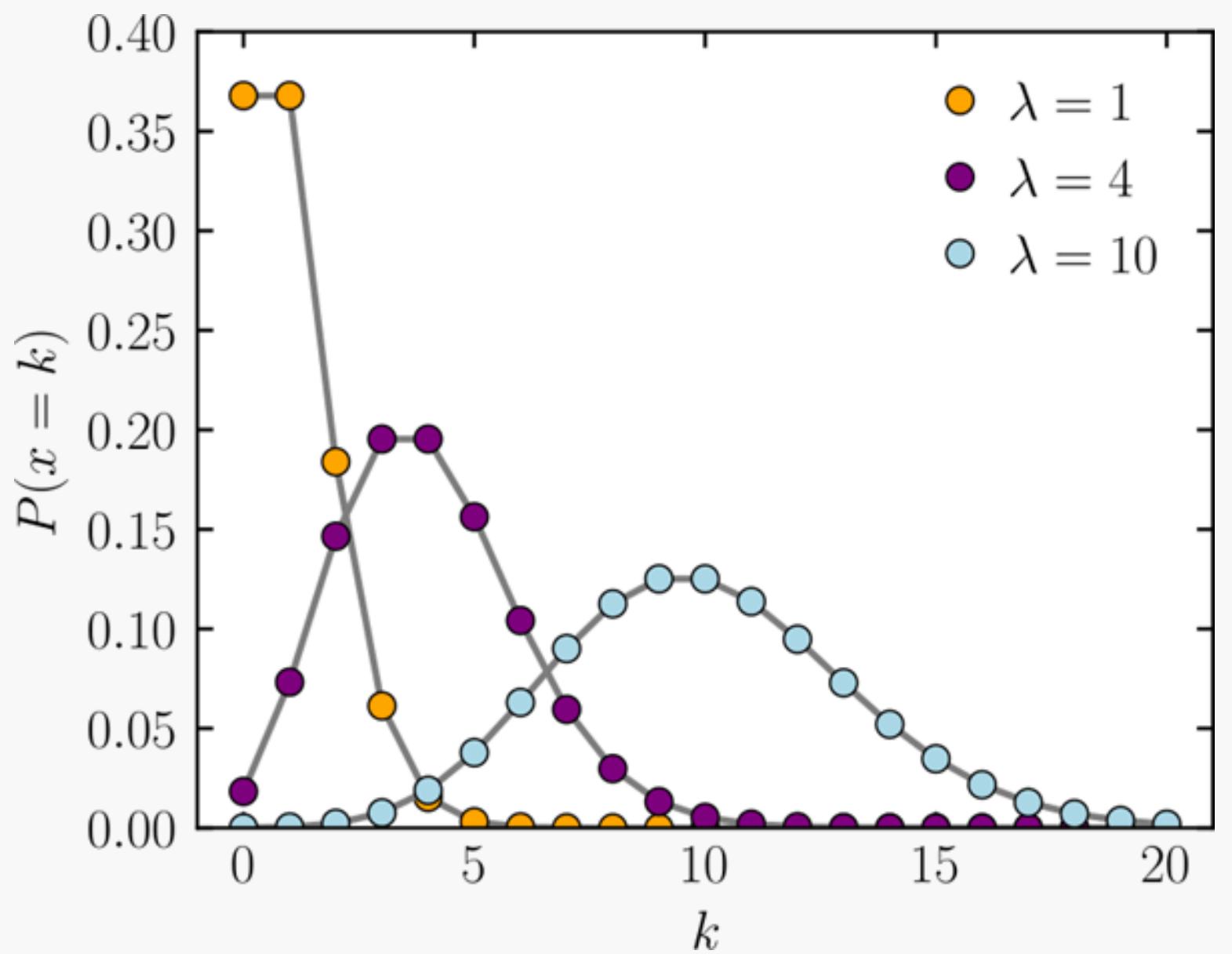
¿Sabes cuantos siniestros te va a dar un asegurado?

¿Sabes que tan graves serán estos siniestros?

NO, y este es el reto, el objetivo es cuantificar el mejor precio posible, basándose en data histórica para estimar el número de siniestros y el costo de cada uno de ellos.

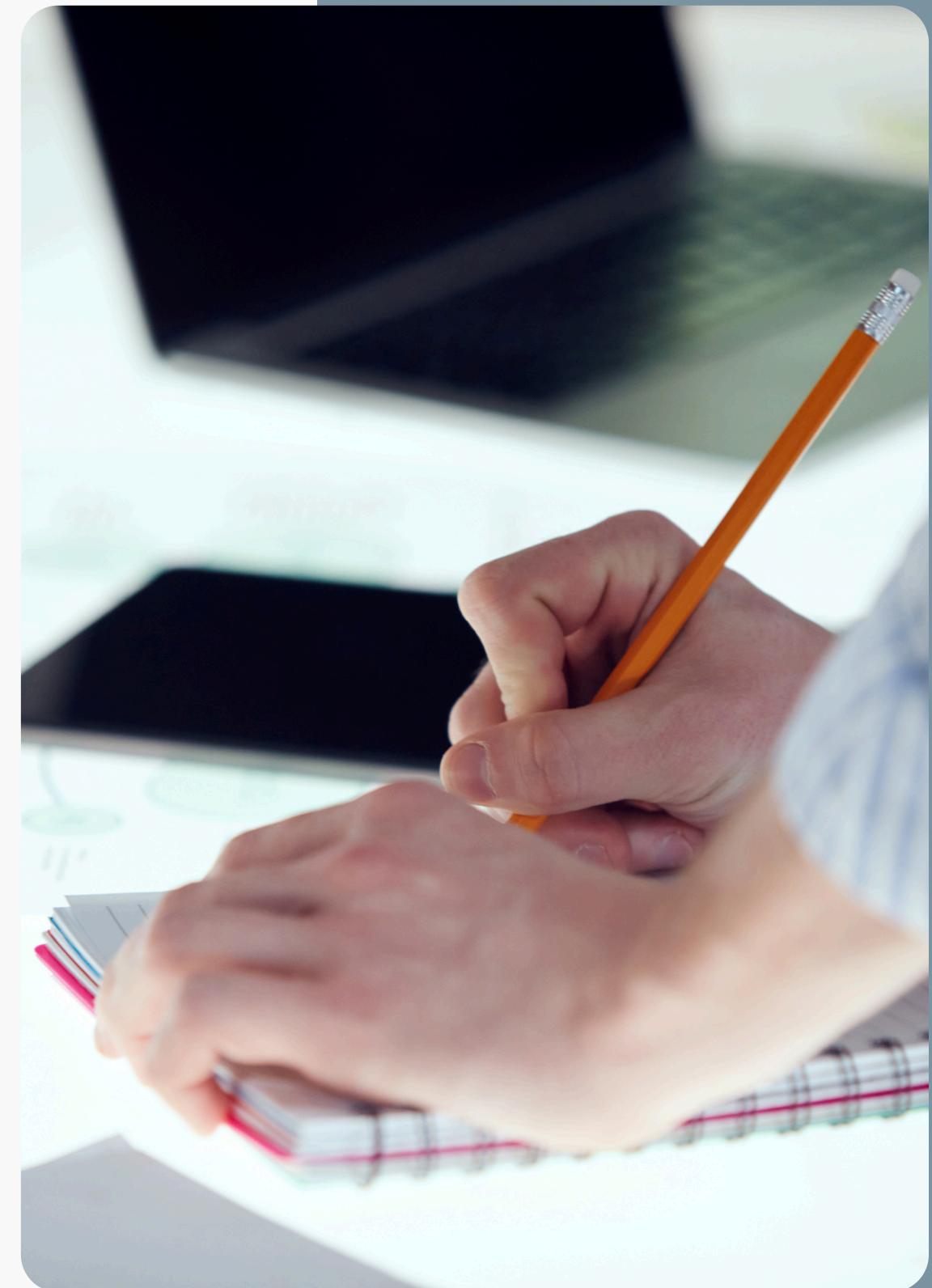
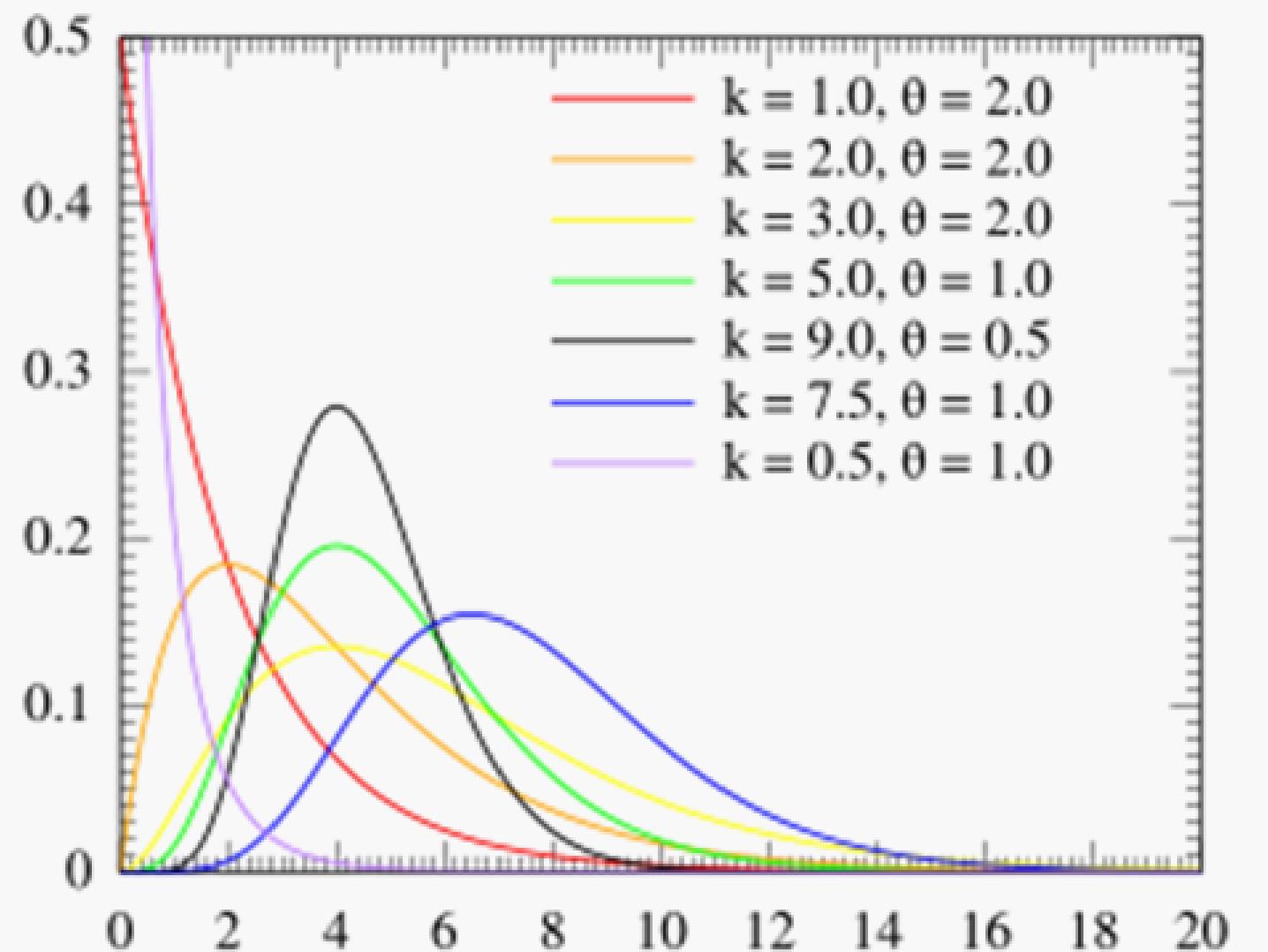
Variables del Pricing

Frecuencia: se modela típicamente con una distribución de Poisson



Variables del Pricing

Severidad: se modela con una distribución Gamma.
(Continua- cola larga)



Precio Mínimo (Sin recargos)

basado en el costo será :

$$P_{perfil} = E[f]_{perfil} * E[S|f]_{perfil}$$

$E[f]_{perfil} \rightarrow$ Frecuencia estimada para un perfil x

$E[S|f]_{perfil} \rightarrow$

Dado que hubo un siniestro que coste tendrá este siniestro para el perfil x

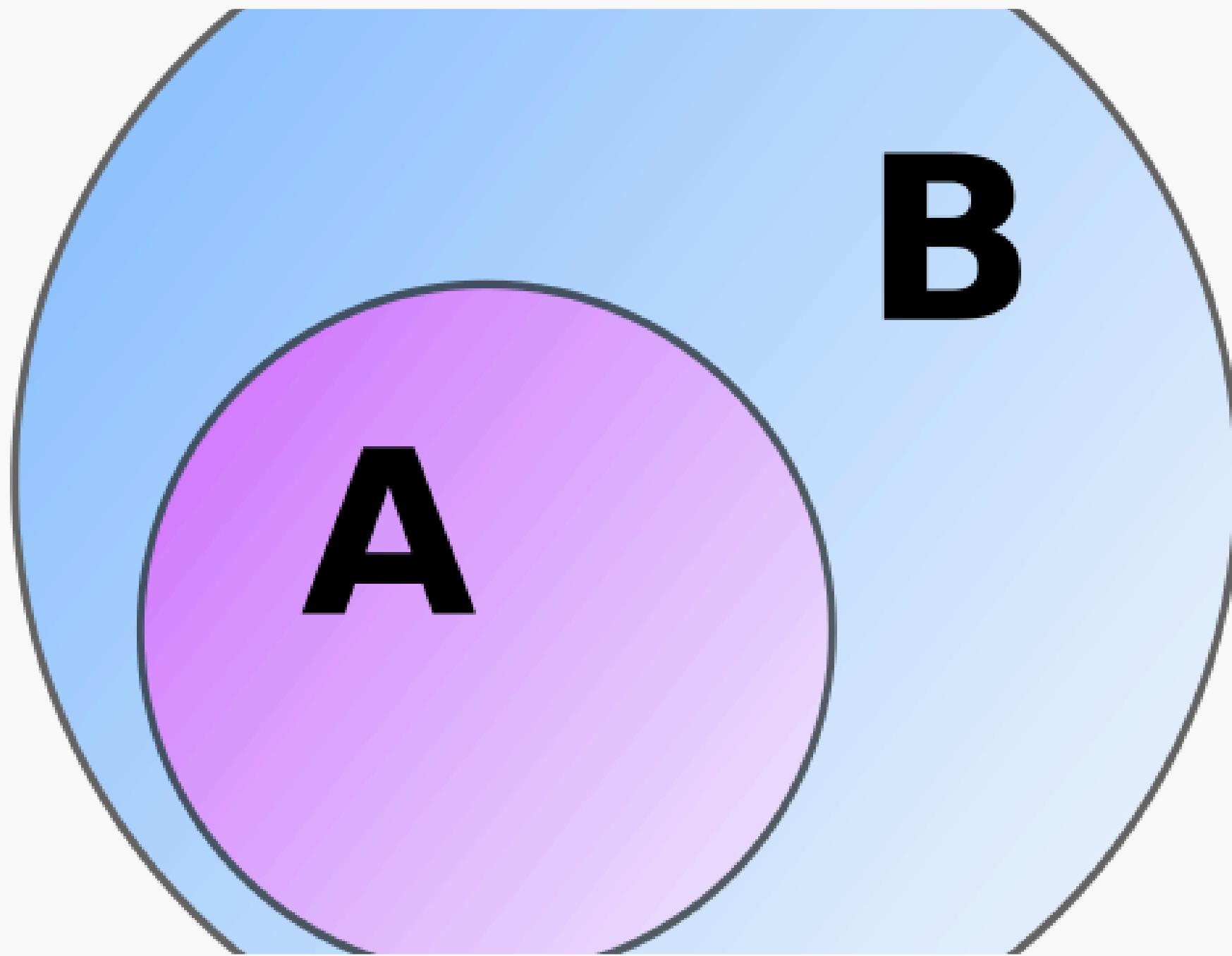
Aspectos legales de la modelización



Dependiendo de la **ubicación geográfica** la regulación es restrictiva en el tipo de modelo que se puede implementar para calcular el precio base o el precio técnico.

En Europa y Latinoamérica la mayoría de las compañías emplean **GLM** como instrumento, en contraste si bien en Estados Unidos la mayoría emplea GLM, otras aseguradoras más modernas, con capacidad técnica alta se han inclinado por complementar la modelación de GLM con otros algoritmos de ML como el **LGBM**, que no es más que otra regresión basada en árboles de decisiones, la cual ha demostrado excelentes resultados, el reto es poder explicar perfectamente a la entidad reguladora el significado de las variables del modelo en la definición del precio.

Aspectos legales de la modelización

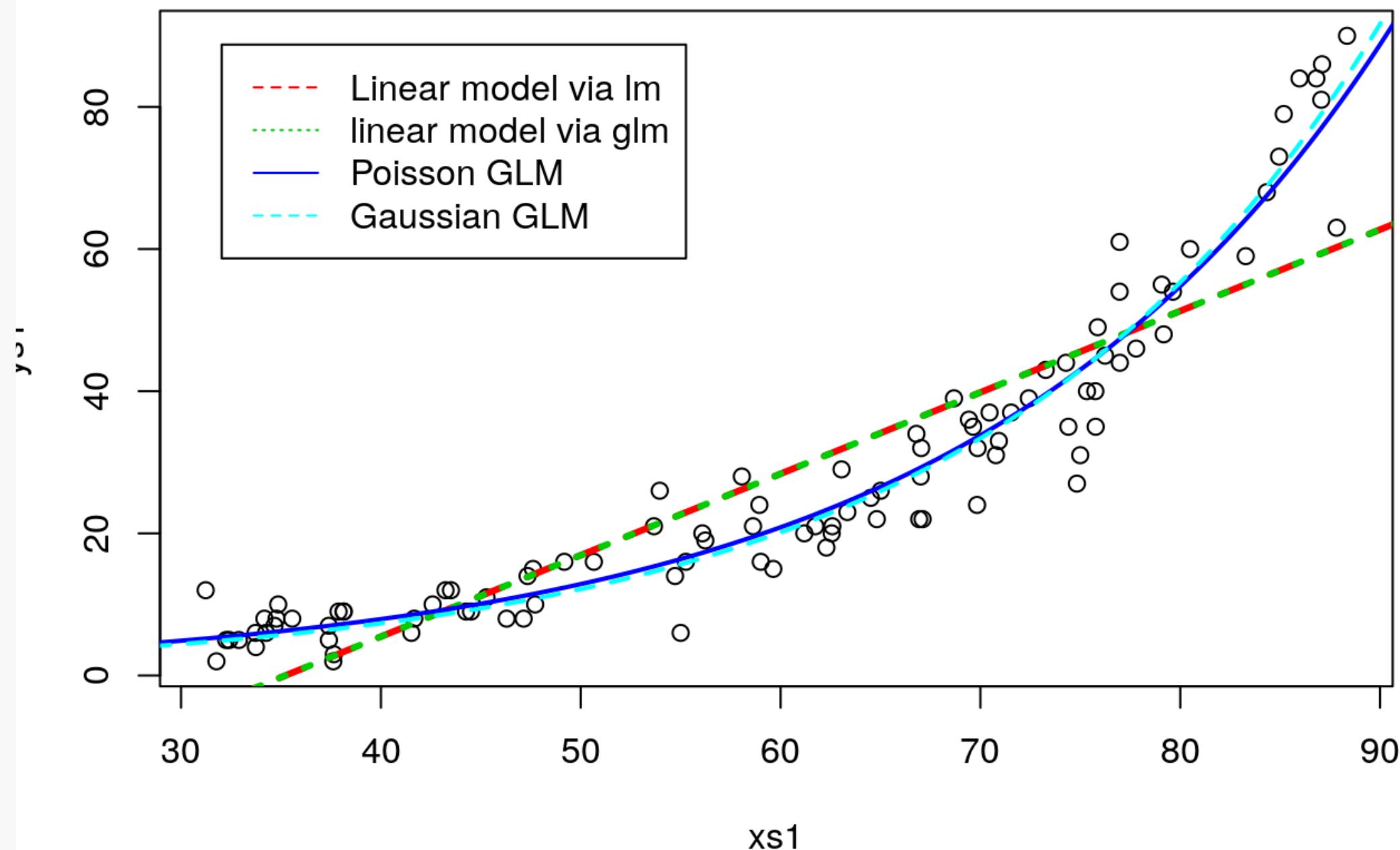


También hay prohibiciones en el tipo de variables a incluir .

- SEXO.
- ORIGEN ETNICO / RAZA.
- RELIGIÓN

GLM (Generalized Linear Model)

¿Por qué un GLM y no una regresión lineal simple?



La **regresión lineal** típicamente se usa para modelar variables respuesta que se comportan como una **distribución normal**, más o menos simétricas respecto a un centro, incluso podrían ser variables negativas. En contraste, un **GLM** permite modelar variables de respuesta con **distribuciones diferentes a la normal**, usualmente distribuciones **sesgadas** que toman valores positivos, como es el caso de la poisson o la gamma.

Aproximación de Poisson: Si λ (el parámetro de Poisson) es grande, la distribución de Poisson se puede aproximar con una normal $N(\lambda, \lambda)$.

Expresión matemática de un GLM

$$g(\mathbb{E}[Y|X]) = X\beta$$

donde:

1. Y es la **variable de respuesta**.
2. X representa la **matriz de predictores** o variables independientes.
3. β es el vector de **parámetros** (incluyendo el intercepto y los coeficientes de cada predictor).
4. $\mathbb{E}[Y|X]$ es la **media condicional** de la variable de respuesta Y , dado el valor de los predictores X .
5. $g(\cdot)$ es la **función de enlace**, que transforma la media condicional para hacerla lineal en los predictores.



$$g(\mathbb{E}[Y|X]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Expresión matemática de un GLM

Recuerde, en el caso de la frecuencia Y se distribuye Poisson y en el caso de la severidad Y se distribuye Gamma.

La función $g(\cdot)$ se conoce como función de enlace, y en el caso de la frecuencia y la severidad $g(\cdot)$ es la función logarítmica.

$$g(\mathbb{E}[Y|X]) = \log(\mathbb{E}[Y|X]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\mathbb{E}[Y|X] = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$$

$$\mathbb{E}[Y|X] = e^{\beta_0} \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \cdot \dots \cdot e^{\beta_p X_p}$$

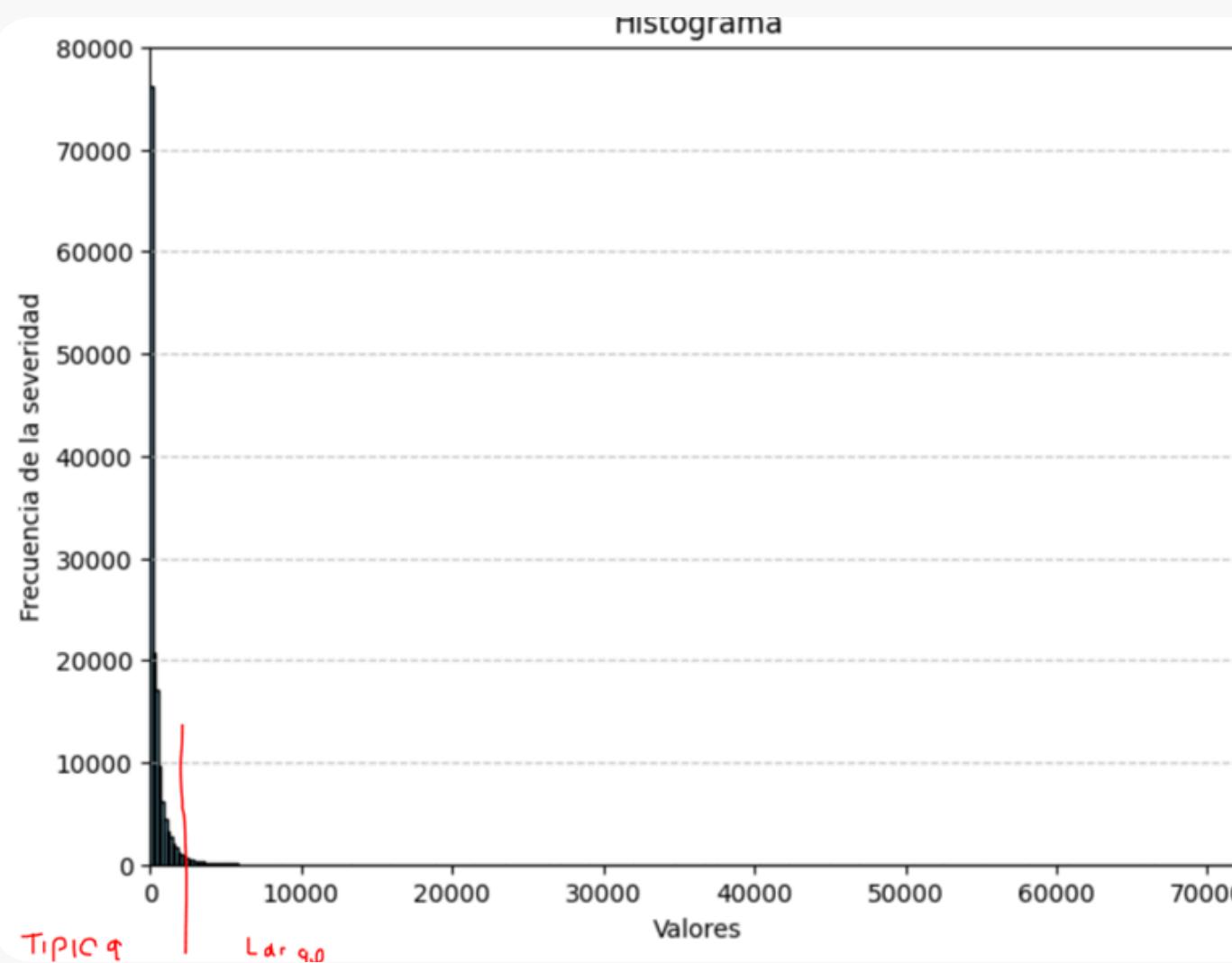
KS test

Hipótesis en el KS Test:

- **Hipótesis nula (H_0):**
"Los datos siguen la distribución teórica especificada (o las dos muestras provienen de la misma distribución)."
 - **Hipótesis alternativa (H_1):**
"Los datos no siguen la distribución teórica especificada (o las dos muestras provienen de distribuciones diferentes)."
-
- $p < 0.05 \rightarrow$ Rechazamos la hipótesis nula \rightarrow Los datos **no siguen** la distribución teórica.
 - $p \geq 0.05 \rightarrow$ No rechazamos la hipótesis nula \rightarrow Los datos **sí podrían seguir** la distribución teórica.

El Kolmogorov-Smirnov Test (KS test) es una prueba estadística no paramétrica que se utiliza para comparar dos distribuciones de probabilidad, o para comparar una distribución de probabilidad teórica con una muestra de datos.

Separación de la severidad típica vs large



Habíamos comentado que la severidad, que se distribuye Gamma es una distribución de cola larga, en ciertos casos es muy muy larga.

◆ Paso 1: Definir una lista de candidatos para x_c

Podemos probar diferentes valores de x_c , por ejemplo, los cuantiles superiores del dataset (ej., percentil 80, 85, 90, etc.).

◆ Paso 2: Ajustar la Gamma y la Pareto

Para cada x_c :

- Ajustamos una distribución Gamma a los datos $x < x_c$.
- Ajustamos una distribución Pareto a los datos $x \geq x_c$.

◆ Paso 3: Realizar el KS test

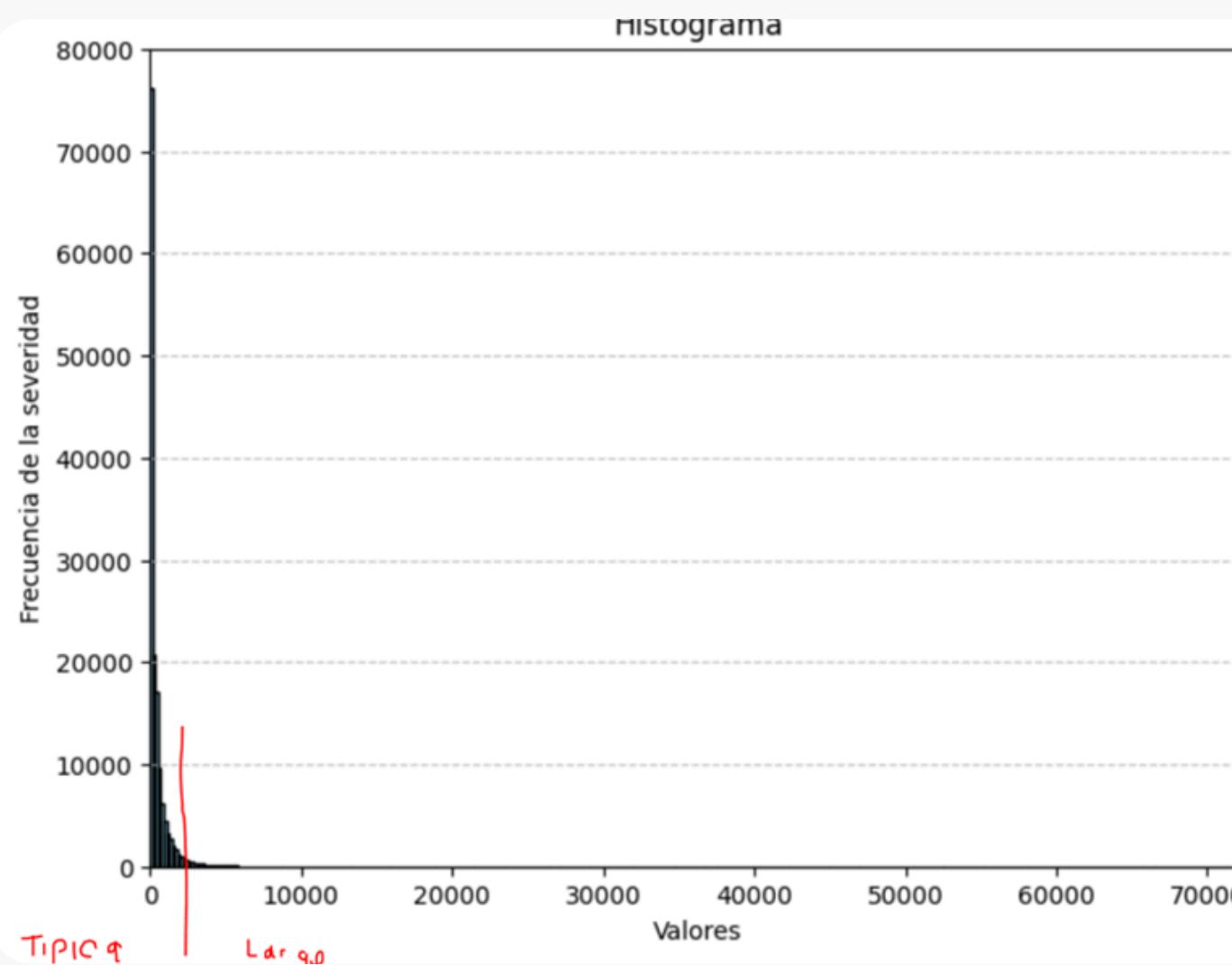
Se aplican dos tests KS:

- KS-Gamma: Compara la parte $x < x_c$ con la Gamma ajustada.
- KS-Pareto: Compara la parte $x \geq x_c$ con la Pareto ajustada.

◆ Paso 4: Elegir el mejor x_c

- Buscamos el mínimo KS test score para ambas distribuciones.
- O elegimos el x_c donde la diferencia entre las dos pruebas KS es mínima.

Separación de la severidad típica vs large



Habíamos comentado que la severidad, que se distribuye Gamma es una distribución de cola larga, en ciertos casos es muy muy larga.

◆ Paso 1: Definir una lista de candidatos para x_c

Podemos probar diferentes valores de x_c , por ejemplo, los cuantiles superiores del dataset (ej., percentil 80, 85, 90, etc.).

◆ Paso 2: Ajustar la Gamma y la Pareto

Para cada x_c :

- Ajustamos una distribución Gamma a los datos $x < x_c$.
- Ajustamos una distribución Pareto a los datos $x \geq x_c$.

◆ Paso 3: Realizar el KS test

Se aplican dos tests KS:

- KS-Gamma: Compara la parte $x < x_c$ con la Gamma ajustada.
- KS-Pareto: Compara la parte $x \geq x_c$ con la Pareto ajustada.

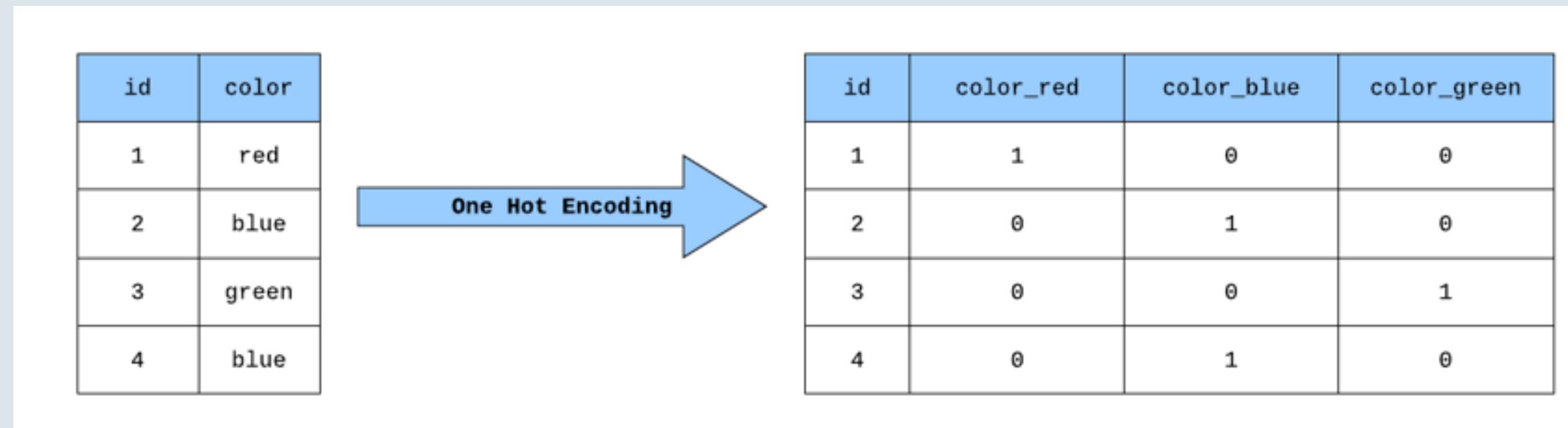
◆ Paso 4: Elegir el mejor x_c

- Buscamos el mínimo KS test score para ambas distribuciones.
- O elegimos el x_c donde la diferencia entre las dos pruebas KS es mínima.

Formatos y variables categóricas

One hot encoder:

- 1.Cuando hay pocas categorías, sino se puede caer en sobre dimensionalidad.
2. Cuando las categorias no tienen un orden coherente (Colores, lugares, ect)



Formatos y variables categóricas

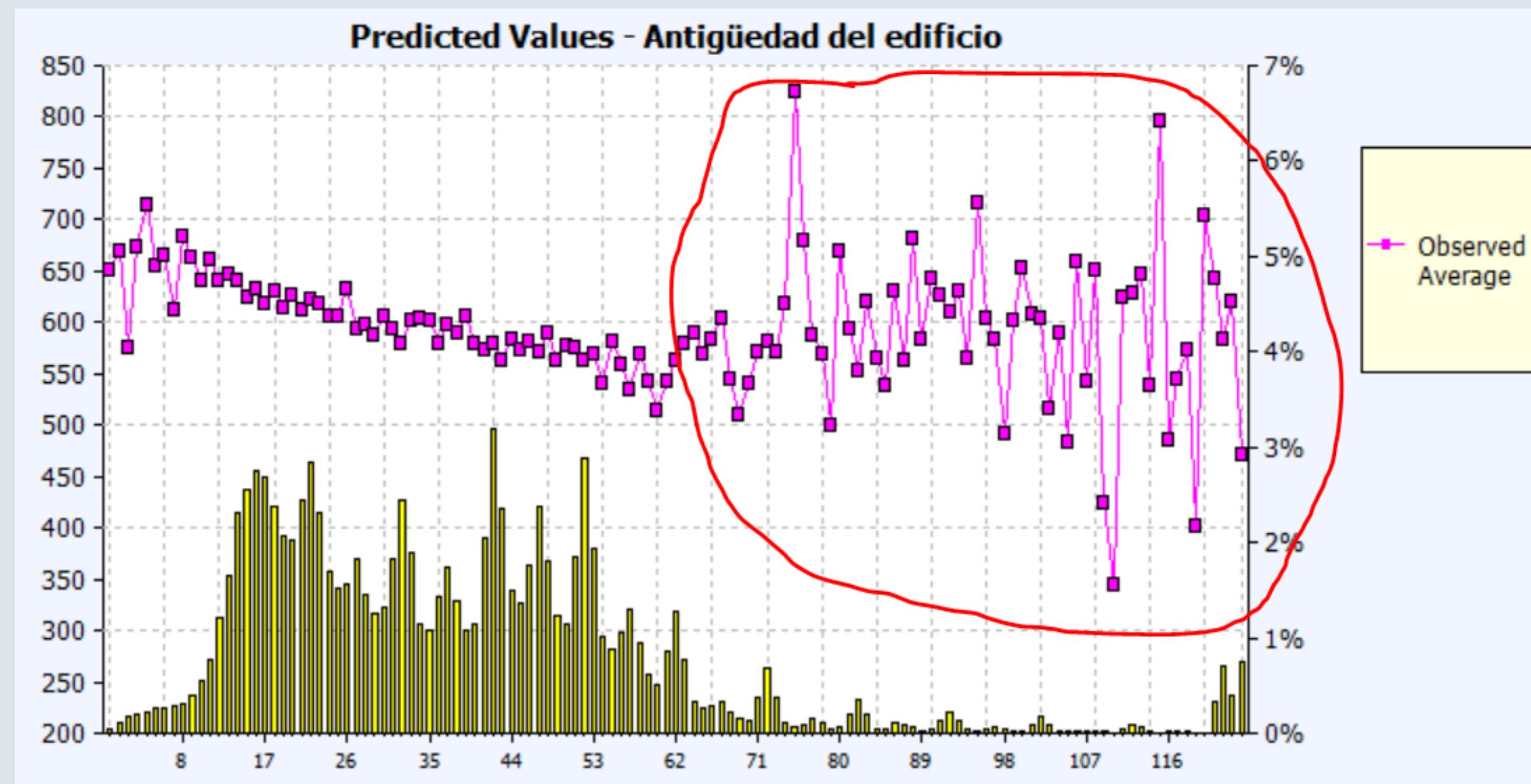
Categorical encoder:

1. Cuando las categorías tienen un orden lógico.
2. Cuando hay muchas categorías.

Customer feedback - 5 point Likert scale	
Feedback	Assign numerical code
Poor	1
Fair	2
Good	3
Very Good	4
Excellent	5

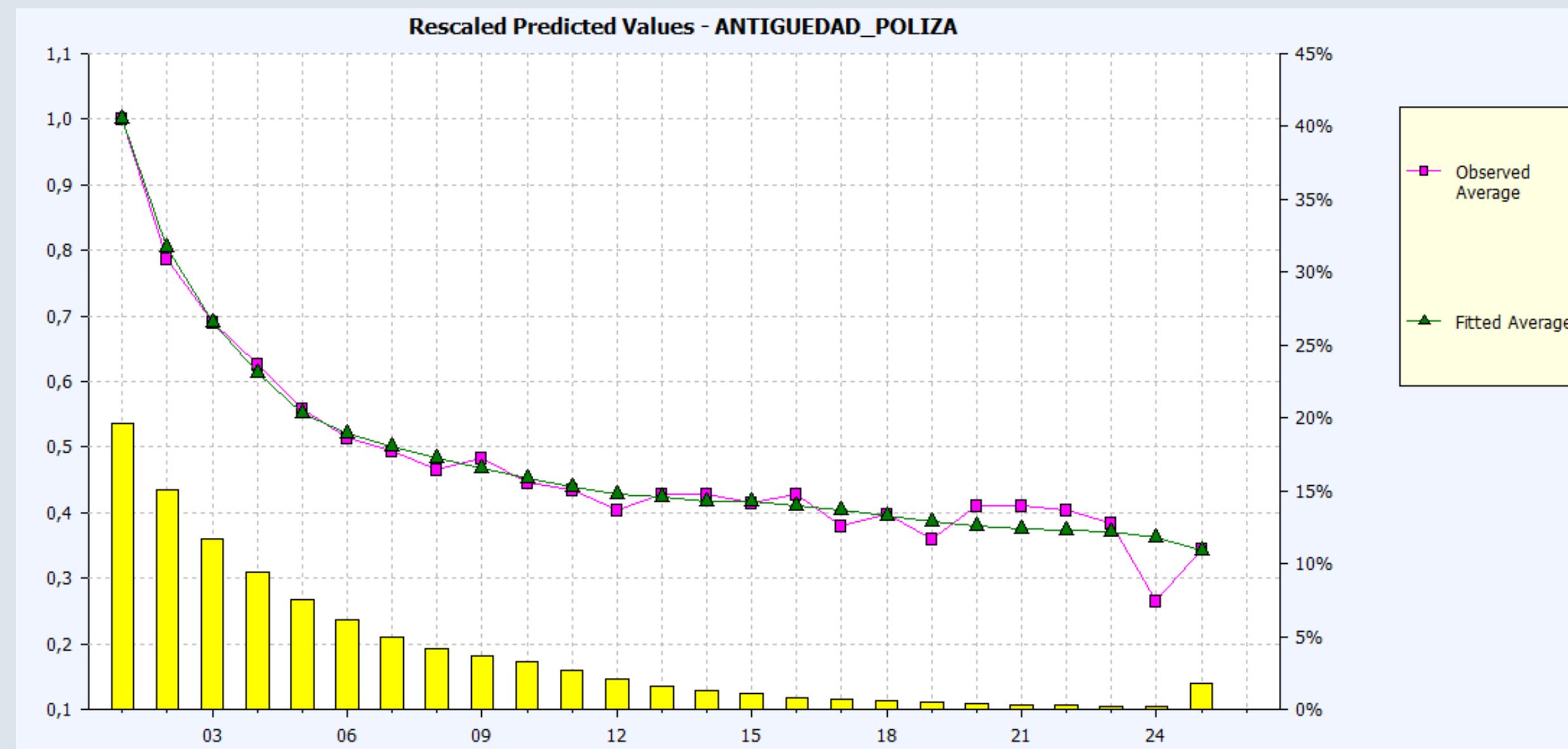
Formatos y variables categóricas

Queremos “Niveles” o categorías que tengan una exposición significativa, a veces puede ser conveniente topar las variables continuas



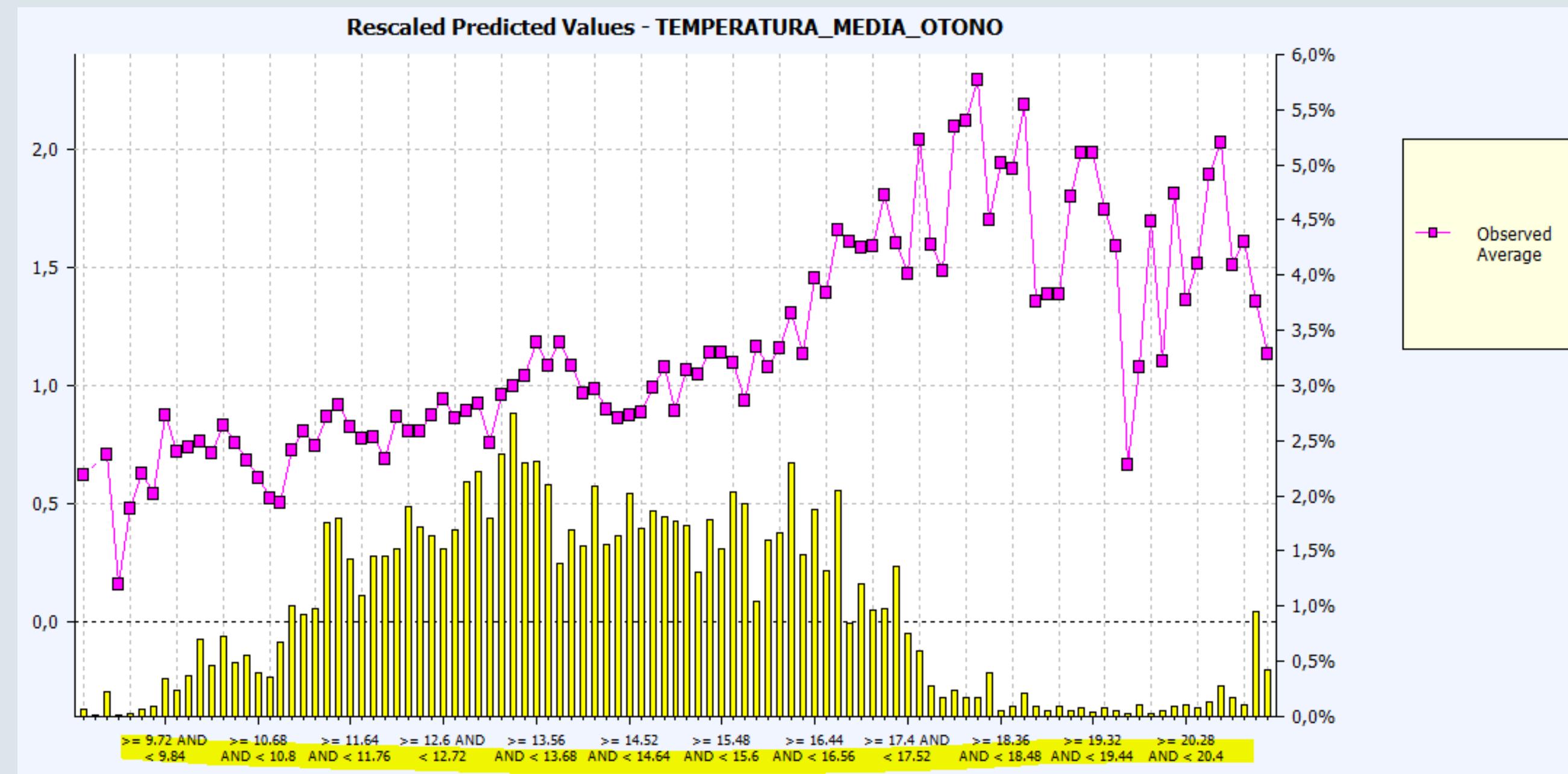
Formatos y variables categóricas

Queremos “Niveles” o categorías que tengan una exposición significativa, a veces puede ser conveniente topar las variables continuas

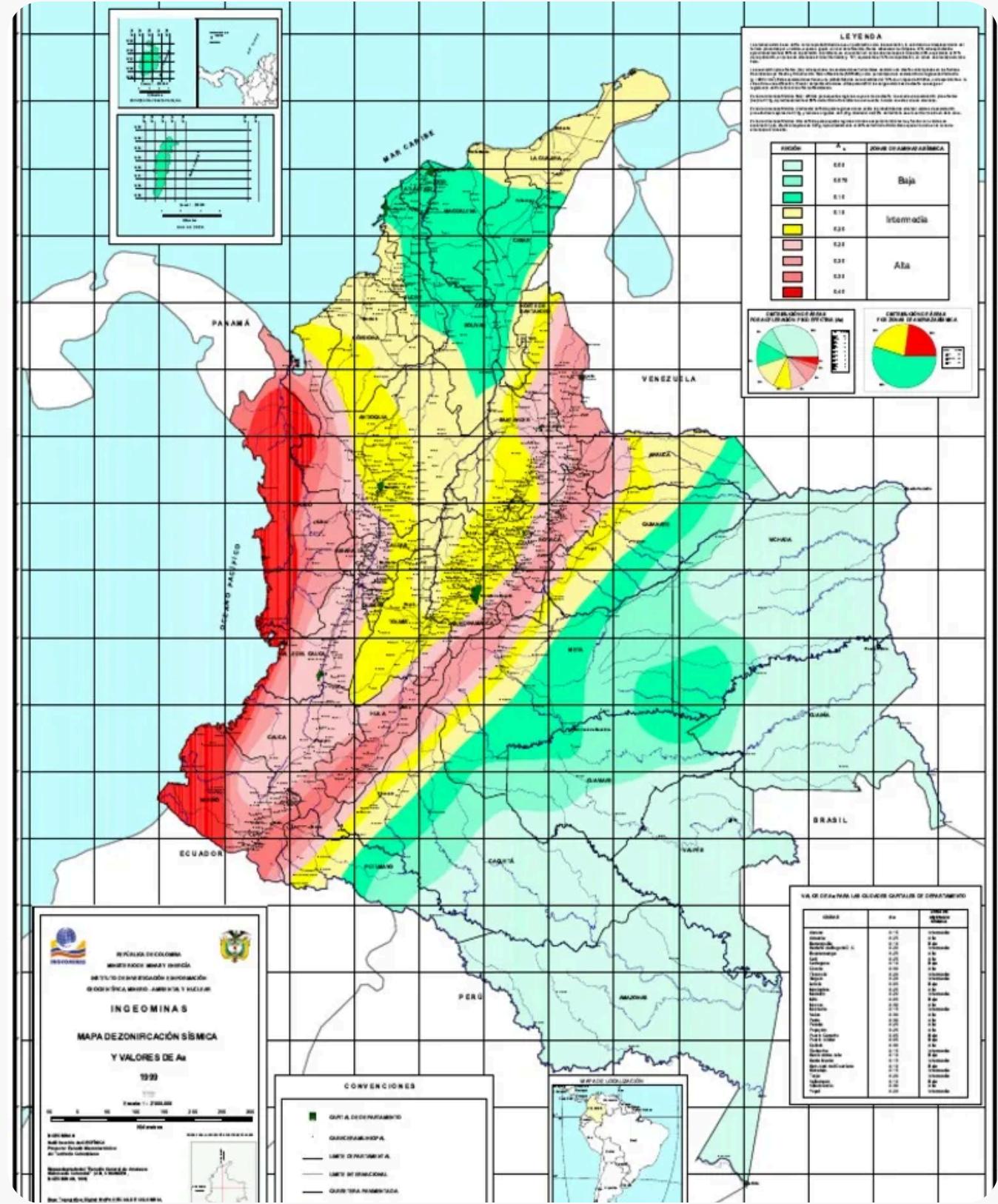


Formatos y variables categóricas

También puede ser conveniente hacer agrupaciones de categorías, por ejemplo intervalos, si las variables son continuas o bien hacer grupos lógicos si las variables son categóricas, por ejemplo agrupar ciudades en municipios.



Reducción de dimensiones

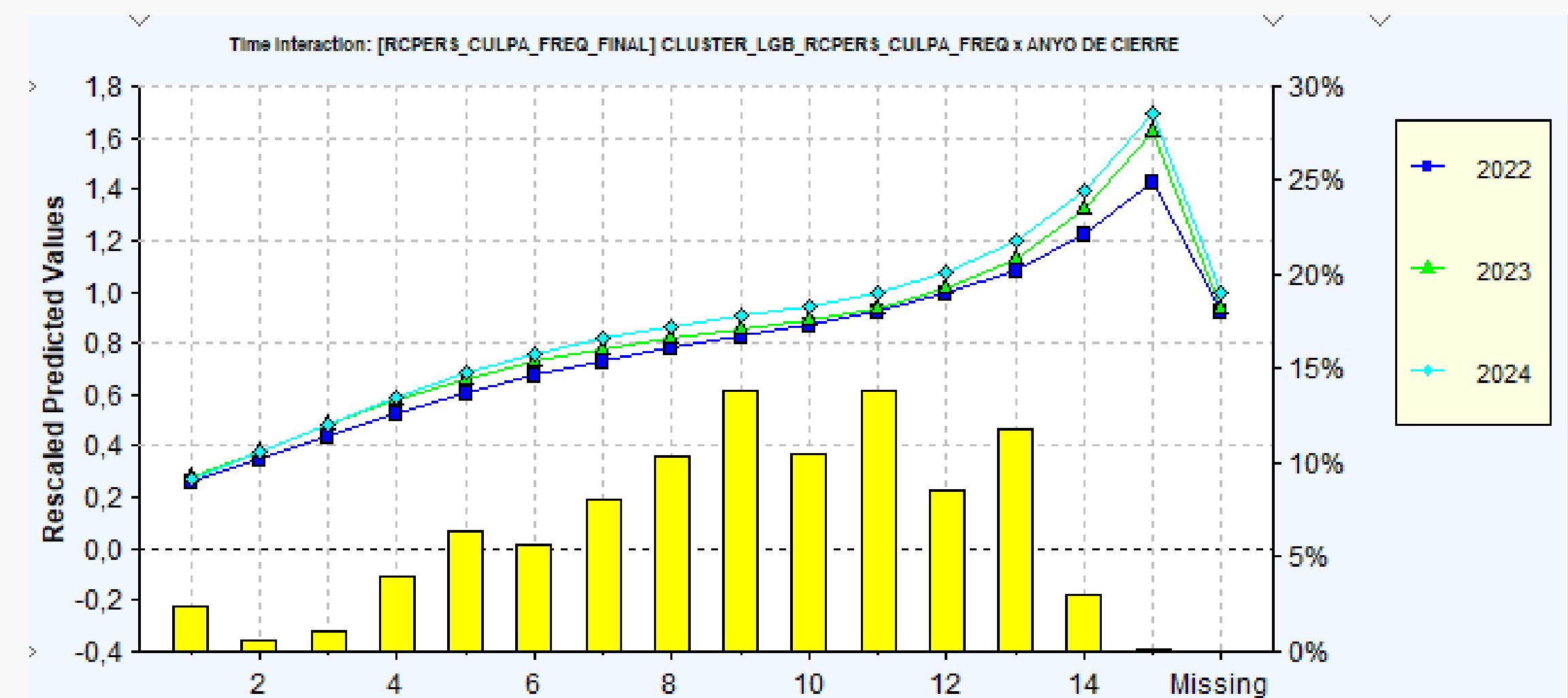


Una entidad bancaria o aseguradora posee mucha información de sus asegurados, entre estos datos de información social y demográfica, típicamente asociados a un código postal u otro nivel de granularidad. Para hacer los modelos una entidad de estas puede contar con fácilmente 400, o 500 variables de este tipo, las cuales en su mayoría están altamente correlacionadas, así, entonces ¿cuál es la mejor variable de todas? ¿Debería usarlas todas?

La microzonificación no es más que una técnica de reducción de dimensionalidad, por medio de algoritmos como el PCA, BYM (Besag York Mollie) entre otras, para “Combinar” todas estas variables en una sola que tenga un mayor efecto predictor para los modelos.

Consistencia temporal

Todas las variables en un modelo deben ser consistentes temporalmente. Esto significa que no ha habido cambio estructural de las variables para el tiempo de recolección de datos



Loadings

Una vez tenemos el precio estimado, este se recarga con ciertos factores para llegar al precio técnico.

$$P_{Tecnico} = P \prod_{i=1}^n Loading_i$$

Loadings: Inflación, Factor de siniestros incurridos, pero no reportados, rentabilidades de grupo, ETC.

Otros análisis



Análisis de competidores, modelos de renovaciones de cartera, anulaciones, inversiones, reaseguro. entre otros.