INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

Desarrollo tecnológico y generación de riqueza sustentable

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de modelación matemática para el desarrollo de planes y proyectos de negocio

4J07 - MODELOS DE PREDICCIÓN EN EMPRESAS Y GOBIERNO MEDIANTE APRENDIZAJE ESTADÍSTICO

Modelación del empleo en el Estado de Jalisco bajo segregaciones sociodemográficas, geográficas y temporales

PRESENTAN

Ingeniero en Finanzas Alicia Karime González Beltrán

Ingeniero en Finanzas Rodrigo Hernández Mota

Ingeniero en Finanzas Raúl Romero Barragán

Profesor PAP: Pablo Dávalos de la Peña

Tlaquepaque, Jalisco diciembre 2017.

ÍNDICE

Contenido

REPORTE PAP - Modelación del empleo en el Estado de Jalisco bajo segregaciones sociodemográficas, geográficas y temporales

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son una modalidad educativa del ITESO en la que el estudiante aplica sus saberes y competencias socio-profesionales para el desarrollo de un proyecto que plantea soluciones a problemas de entornos reales. Su espíritu está dirigido para que el estudiante ejerza su profesión mediante una perspectiva ética y socialmente responsable.

A través de las actividades realizadas en el PAP, se acreditan el servicio social y la opción terminal. Así, en este reporte se documentan las actividades que tuvieron lugar durante el desarrollo del proyecto, sus incidencias en el entorno, y las reflexiones y aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

La correcta medición de indicadores es crucial para conocer el desarrollo de una entidad y el cambio del mismo a través del tiempo. Uno de los indicadores claves para poder saber el comportamiento de una región es el dato del empleo (trabajadores asegurados), dato que por sí solo da certeza sobre la estabilidad pero además es un excelente vía de información para poder detectar comportamientos atípicos y potencialmente dañinos para la región en cuestión.

Es por lo anteriormente mencionado, que crece una necesidad de contar con herramientas que con un alto grado de precisión, puedan dar una idea sobre el comportamiento del empleo en distintos niveles de segregación.

1. Introducción

El empleo es un dato de suma importancia en distintos niveles que inician en instancias municipales pero llegan a ser de comparación incluso internacional. Es además el insumo básico de los ciudadanos para poder financiar y mantener un cierto estilo de vida, siempre en función de lo que puedan obtener a través de un empleo formal.

En adición a lo anteriormente mencionado, se reconoce que el empleo es un dato crucial para que la población pueda hacer una comparación de calidad de vida entre distintas regiones, por lo que el dato del empleo se vuelve algo crucial de conocer no solo para que las autoridades puedan saber el rumbo de cierta región e identificar posibles cánceres, sino también para que la población pueda entender qué está pasando y dónde puede obtener un mayor nivel de vida.

1.1. Objetivos

1. Entender el comportamiento de las series para poder detectar qué variables influyen en el empleo.
2. Poder comprender, con ayuda de las variables más importantes detectadas durante el análisis, la segregación que más conviene para poder entender y estimar el dato de empleo en distintas zonas y/o niveles representativos.
3. Tener un abanico de propuestas para poder tomar una decisión informada en base a distintas métricas de desempeño.
4. Tener lista una o varias propuestas replicables para poder incorporar datos, aprender de los mismos y arrojar resultados interpretables para el Instituto de Información Estadística y Geográfica del Estado de Jalisco sobre el empleo en la entidad.

1.2. Justificación

Los datos representan una abstracción de cierta situación y contexto. La mala interpretación de los mismos puede tener graves consecuencias respecto a la toma de decisiones o evaluación de impacto en proyectos de desarrollo. En el caso particular del empleo, entender la dinámica y comportamiento de esta variable se vuelve obligatorio para medir efectos de políticas públicas y decisiones gubernamentales. Así mismo, generar proyecciones puede determinar si es prudente comenzar cierto programa social además de ayudar a determinar el estado de "productividad" en cierta región.

En este proyecto abordamos el tema de pronósticos respecto a la variable de empleos asegurados en el estado de Jalisco, atendiendo distintos niveles de segregación; desde municipal hasta división económica y variables sociodemográficas (edad y género).

1.3 Antecedentes

El IIEG (Instituto de Información Estadística y Geográfica) se especializa en el análisis e interpretación de datos socioeconómicos a nivel Jalisco para fortalecer el desarrollo del estado y contribuir en la toma de decisiones gubernamentales con un fuerte sustento estadístico.

En puntos anteriores del tiempo, el IIEG (Instituto de Información Estadística y Geográfica del Estado de Jalisco) se dedicó a hacer estudios con altos niveles de precisión que rondaban el 10% de error aproximádamente.

1.4. Contexto

En IIEG se ha trabajado previamente en pronosticar la serie de tiempo de empleos asegurados. Los datos disponibles se encuentran en un "cubo de información" llamado COGNOS (gestionado por IBM). La base de datos no se actualiza en tiempo real y por lo tanto la información suele estar rezagada ciertos meses.

El método de estimación que utilizan genera un error aproximadamente del 10% mediante regresión lineal convencional. Estos métodos son a un nivel agregado (total estatal). Se tiene interés en desarrollar un modelo que pueda estimar a varios periodos en el futuro y bajo diferentes variables de segregación (municipal, económicas y sociodemográficas).

La zona metropolitana de Guadalajara está compuesta por los siguientes municipios:

1. Guadalajara

2. Zapopan
3. Tonalá
4. Tlaquepaque
5. Tlajomulco de Zúñiga
6. El Salto
7. Ixtlahuacán de los Membrillos
8. Zapotlanejo
9. Juanacatlán

Además de considerar como parte del estudio (y cada que se menciona AMG) a:

1. Lagos de Moreno
2. Puerto Vallarta

2. Desarrollo

2.1. Sustento teórico y metodológico

Dentro del proyecto actual se utilizan distintas herramientas para poder realizar las estimaciones necesarias, herramientas que en las secciones venideras se describen a mayor detalle.

2.1.1 Métricas de medición del error

Todos los modelos aquí descritos usaron una o más de las métricas aquí expuestas. Por motivos de espacio y concretización de ideas, se describen una única vez a continuación, siendo únicamente referenciadas posteriormente por su nombre.

2.1.2.1 Error medio absoluto

Este error es la media de la diferencia entre entre el valor real y la estimación del mismo en términos absolutos para evitar la afectación de números negativos y positivos, dado por la ecuación 1.

Ecuación 1: Error medio absoluto

$$\mathrm{MAE}(y, \hat{y}), = \frac{1}{N} \sum_{i=0}^{N} |y_i - \hat{y}_i|$$

2.1.2.2 Error mediano absoluto

Es la mediana de la diferencia entre el dato real y la estimación del mismo. A diferencia del error medio absoluto, permite analizar la diferencia entre los datos quitando la afectación de posibles outliers[1].

Ecuación 2: Error mediano absoluto

$$\text{MedianAE}(y, \hat{y}), = \text{median}\left(|y_1 - \hat{y}_1|, \cdots, |y_n - \hat{y}_n|\right)$$

2.1.2.3 Raíz del error cuadrático medio (RMSE).

La raíz del error cuadrático medio se usa para realizar una comparación entre dos conjuntos de datos, el estimador y el valor real que se quiere calcular, con la característica de que la serie debe tener una separación homogénea en el tiempo y ambas series tienen que tener el mismo tamaño.Se define de manera algebraica en la ecuación 3.

Ecuación 3: Raíz del error cuadrático medio

$$\text{RMSE} = \sqrt{\frac{\text{SSE}}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

2.1.2.4 Error porcentual

Error entre la diferencia de las estimaciones en un conjunto de tamaño *n* y el mismo conjunto, en el mismo tiempo y de mismo tamaño de los datos reales. Dado por la ecuación 4.

Ecuación 4: Error porcentual

$$perror = \frac{\sum_{i=0}^{n} |y_i - \hat{y_i}|}{\sum_{i=0}^{n} |y_i|}$$

[1] Se le dice así a los datos atípicos.

2.1.2 Modelación econométrica

2.1.2.1 Modelos econométricos

2.1.2.1.1 Modelo Autorregresivo (AR)

Un modelo autorregresivo es una representación de un tipo de proceso aleatorio, que como tal, describe ciertos procesos variables en el tiempo. Especifica que la variable de salida depende linealmente de sus propios valores anteriores.

La notación *AR(p)* presenta un modelo autorregresivo de orden *p*. Se define de la siguiente manera:

Ecuación 5: Modelo autorregresivo

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t$$

2.1.2.1.2 Modelo de Medias Móviles(MA)

Un modelo de medias móviles constituye un tipo de estructura estocástica lineal, los valores que toma la variable dependiente se explican por los efectos de los *shocks* aleatorios que se hayan producido en el momento *t, t-1,t-2...t-p.*

Tienen como característica ser siempre estacionarios.

Ecuación 6: Modelo de medias móviles

$$X_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}$$

2.1.2.1.3 Modelo Autorregresivo de Medias Móviles (ARMA)

Un modelo autorregresivo de medias móviles es un modelo estadístico que está formado por dos partes, una parte autorregresiva *AR(p),* y otra de media móvil *MA(q)* que presentan un componente de estacionalidad.

Se utiliza para entender y predecir futuros valores de la serie, se aplica a series temporales de datos.

Ecuación 7: Modelo Autorregresivo de Medias Móviles

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + w_{t-q}$$

2.1.2.1.3 Modelo Autorregresivo Integrado de Medias Móviles (ARIMA)

Un modelo autorregresivo integrado de medias móviles es un modelo estadístico que maneja variaciones y regresiones de datos con el objetivo de encontrar patrones en la serie y así poder realizar una predicción futura. Es un modelo dinámico de series temporales donde las estimaciones se explican por datos históricos.

Este modelo presenta una gran sensibilidad a la precisión con la que se seleccionen los coeficientes.

Ecuación 8: Modelo Autorregresivo Integrado de Medias Móviles

$$(1 - \beta_1 L^1 - \cdots - \beta_k L^k)\nabla^d Y_t = (1 - \theta_1 L^1 - \cdots - \theta_k L^k)\upsilon_t$$

2.1.2.1.3 Modelo Autorregresivo Integrado de Medias Móviles Estacional (SARIMA)

Un modelo autorregresivo integrado de medias móviles con estacionalidad es un modelo que presenta las mismas características que un modelo ARIMA pero agregándole la parte estacional, que usualmente tiene patrones de manera mensual, anual o trimestral.

Ecuación 9: Modelo Autorregresivo Integrado de Medias Móviles Estacional

$$(1 - \beta_1 L^1 - \cdots - \beta_k L^k)(1 - \psi L^s)\nabla^{d,s} Y_t$$

2.1.2.2 Metodología Box Jenkins

La metodología de *Box-Jenkins*, en el análisis de series de tiempo, se aplica a los Modelos Autorregresivos de Media Móvil (ARMA) o los Modelos Autorregresivos Integrados de Media Móvil (ARIMA) para encontrar el mejor ajuste de una serie temporal de valores, a fin de que los pronósticos sean más acertados.

Es un proceso iterativo que consiste en:

- Identificación del modelo
    - Analizar que la serie sea estacionaria.
    - Identificación de la estacionalidad de la serie dependiente.
    - Uso de autocorrelación y autocorrelación parcial de la serie de tiempo para la elección de rezagos.
- Estimación de parámetros.
    - Algoritmos de cálculo para obtener coeficientes que mejor se ajusten al modelo seleccionado.
        - Máxima verosimilitud

        - Mínimos cuadrados no lineales
- Verificación del modelo, si el modelo estimado se ajusta a las especificaciones de un proceso univariado estacionario.
    - Los residuales deben ser independientes el uno del otro.
    - La media de los residuales debe ser constante en el tiempo.
    - La varianza de los residuales debe ser constante en el tiempo.

2.1.3 Modelación *machine learning*

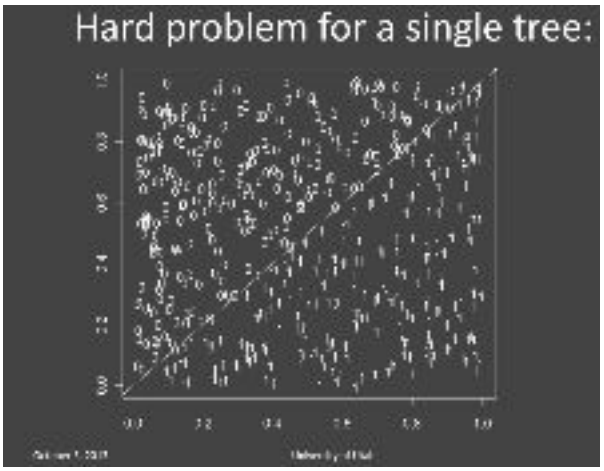
2.1.3.1 Algoritmos de machine learning

2.1.3.1.1 Random Forest

Random forest es un algoritmo que entrena a *n* árboles de decisión en sub-muestras de los datos de entrenamiento y usa el promedio de la estimación de cada árbol para evitar el *overfitting*[2] y mejorar la capacidad de predicción.

El siguiente problema muestra gráficamente la flexibilidad que ofrece el algoritmo de *random forest* en comparación con solamente 1 árbol de decisión.

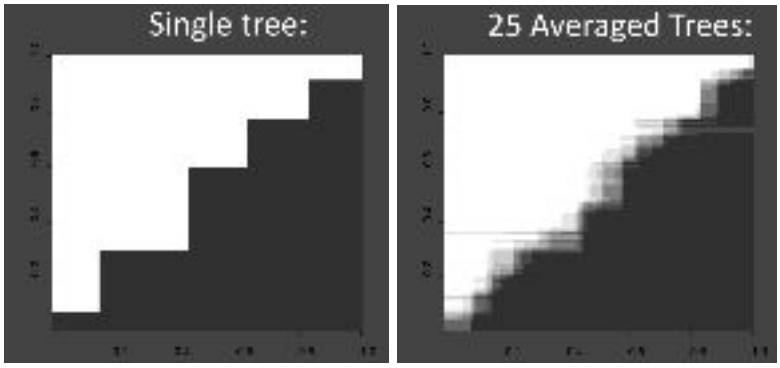
Figura 1: Flexibilidad de un random forest respecto a un árbol de decisión



Fuente: USU

[2] Sobre ajuste de los datos.

Figura 2: Flexibilidad de un random forest respecto a un árbol de decisión



Fuente: USU

2.1.3.1.2 Gradient Tree Boosting

Es una implementación de los algoritmos de boosting en donde se especifica una función diferenciable de costo arbitraria. Destacado por:

1. Manejo natural de diferentes tipos de variables.
2. Alto poder predictivo.
3. Robustez contra *outliers*.

2.1.3.1.3 Support Vector Machines

Support Vector Machines es un algoritmo que destaca por:

1. Efectividad en espacios de alta dimensionalidad
2. Puede ser efectivo en casos donde el número de dimensiones es mayor que el número de muestras.
3. Versatilidad, debido a la posibilidad de adaptación de distintos *kernels* que, en términos simples, es la transformación de los datos mediante diferentes técnicas.

La función de costo que se establece para el algoritmo es la siguiente

Ecuación 10: Función de costo SVR

$$C\sum_{i=0}^{n}(\xi_i+\xi_i^*)+\frac{|w|^2}{2}$$

Este algoritmo se basa en la idea de encontrar un hiperplano que divida de mejor manera (en el caso de clasificación) las dos clases, siendo los vectores soporte los puntos que mejor *soportan* o crean la frontera.

2.1.3.1.4 Lasso

Lasso es un modelo lineal conocido por su tendencia a identificar coeficientes más relacionados con la variable de respuesta reduciendo la cantidad de coeficientes hasta aquellos donde la solución dada sea dependiente. De forma concreta, se puede identificar como una regresión clásica pero con penalización en L1, definiendo la penalización en L1 como:

Ecuación 11: Función de costo clásica L1

$$S = \sum_{i=1}^{n} |y_i - f(x_i)|.$$

Cabe mencionar, que se aplica el concepto anterior específicamente a los pesos, teniendo definida una función de costo de la siguiente manera:

Ecuación 12: Función de costo Lasso

$$\min_{w} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1$$

Lasso entonces, minimiza mínimos cuadrados (como una regresión clásica) con el *plus* de la segunda parte de la ecuación, la norma en L1 de los pesos multiplicada por una constante.

2.1.3.1.5 Ridge

Ridge, al igual que Lasso, es de igual manera una regresión teniendo la diferencia de que el término específico de regularización es, sobre los pesos pero utilizando la norma en L2 y asumiendo la siguiente función de costo:

Ecuación 12: Función de costo Ridge

$$\min_{w} ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$$

El término alfa ha controla la cantidad de encogimiento, relacionando la magnitud con la robustez a la colinealidad.

2.1.3.1.6 Técnica ponderada de estimación

La técnica ponderada de estimación es una propuesta en este proyecto para poder obtener estimaciones mensuales a partir de una estimación anual (véase 2.2.1.3).

Dicha técnica toma una matriz ordenada de datos reales a través del tiempo como la siguiente[3]:

Tabla 1: Datos de ejemplo distribuidos por mes

	Enero	Febrero	Marzo
2010	10	70	100
2011	11	71	100

Fuente: Elaboración propia

Y la transforma en una como la siguiente:

Tabla 2: Ponderaciones de ejemplo aplicadas

	Enero	Febrero
2010	.10	.70
2011	.11	.71

Fuente: Elaboración propia

El propósito es tomar como mes referencia un mes (en base al cual se hacen las proyecciones anuales, de *mes elegido* a *mes elegido.* Tomada una referencia, se obtiene qué proporción representan los demás componentes (meses) del año en cuestión, tal y como se ve en la tabla 2.

La ponderación final aplicada es un promedio por columna, para tener como referencia el cambio a través del tiempo de la representación de las ponderaciones por mes, siendo este caso 0.105 para enero y 0.755 para febrero.

2.2. Planeación y seguimiento del proyecto

[3] Se asumen un periodo de un ano compuesto por tan solo 3 meses a manera de simplificación.

2.2.1 Descripción del proyecto

El proyecto se basa en encontrar la forma adecuado de modelación para periodos posteriores que mejor se ajuste a los datos así como la comparación entre una métrica clásica de evaluación como lo es una metodología clásica (la econométrica) y una nueva propuesta como lo es el *machine learning* atacando dos problemas: el estatal y el municipal.

Se describe a continuación el procedimiento de cada una de las metodologías propuesta.

2.2.1.1 Modelo econométrico

La propuesta econométrica se realizó a nivel municipal, específicamente aplicado al área metropolitana de Guadalajara. Se utilizó la serie de tiempo de empleo de cada uno de los municipios tomando datos mensuales de diciembre 2000 hasta agosto de 2017.

El código realizado sigue la metodología Box Jenkins (descrita en la sección 2.2) y es capaz de ajustar un modelo econométrico, de acuerdo a las características de la serie, *AR(p), MA(q), ARMA(p,q), ARIMA(p,d,q) o bien SARIMA(p,d,q)(P,D,Q,S).*

Primeramente, se realiza un análisis de la serie original,en la cual se analiza la estacionariedad a través de una prueba *Dickey Fuller Aumentada*, de la misma manera se analiza la estacionalidad de la serie evaluada.

Una vez realizada dicha prueba, se realizan a la serie diferencias para, en caso de no ser estacionaria, se transforme en una serie estacionaria que sea adecuada para el ajuste del modelo.

Cuando la serie ya es estacionaria, se seleccionan parámetros iniciales para ajustar un primer modelo a través de la autocorrelación y autocorrelación parcial. Se genera entonces un primer modelo, del cual, se evalúa la significancia de sus parámetros para mejorar el ajuste del modelo y crear un segundo modelo con todos los parámetros significativos.

Cuando se obtiene el modelo adecuado, se realiza una predicción del empleo de diciembre de 2015 a agosto de 2017, esto para poder realizar un análisis del error que tuvo cada modelo realizado para cada municipio.

Finalmente se analiza el error de predicción de cada modelo seleccionado, con las métricas de error mencionadas anteriormente, *RMSE, MedianAE, MeanAE* y el error porcentual.

El proceso tiene como objetivo obtener una estimación final del empleo a 2018 de manera mensual a través de una vista clásica, y servir de *benchmark*[4] contra los modelos de *machine learning* propuestos en secciones posteriores.

2.2.1.2 Modelo estatal

La propuesta de predicción estatal se basa en considerar diferentes niveles de segregación sobre la variable de respuesta. Se considera que es posible aumentar el nivel predictivo del modelo usando datos que proporcionan cierto nivel de detalle. En este sentido, el modelo podría diferenciar patrones que se presentan únicamente en cada nivel de segregación. Podemos clasificar las variables de segregación de la siguiente forma:

1. Variables Economicas
    a. División económica
        i. Representa los distintos sectores económicos a nivel Jalisco. Puede adquirir los siguientes valores:
            1. Agricultura, ganadería, silvicultura, pesca y caza
            2. Comercio
            3. Industria de la construcción
            4. Industria eléctrica, captación y suministro de agua potable
            5. Industrias de transformación
            6. Industrias extractivas
            7. Servicios
            8. Transportes y comunicaciones
2. Variables Sociodemográficas

[4] Punto de comparación

a. Rango de edad

Representa el rango de edad de la población. Puede adquirir los siguientes valores:

i. De 15 a 19 años.
ii. De 20 a 24 años.
iii. De 25 a 29 años.
iv. De 30 a 34 años.
v. De 35 a 39 años.
vi. De 40 a 44 años.
vii. De 45 a 49 años.
viii. De 50 a 54 años.
ix. De 55 a 59 años.
x. De 60 a 64 años.
xi. De 65 a 69 años.
xii. De 70 a 74 años.
xiii. De 75 ó más años
xiv. Menor a 15 años.

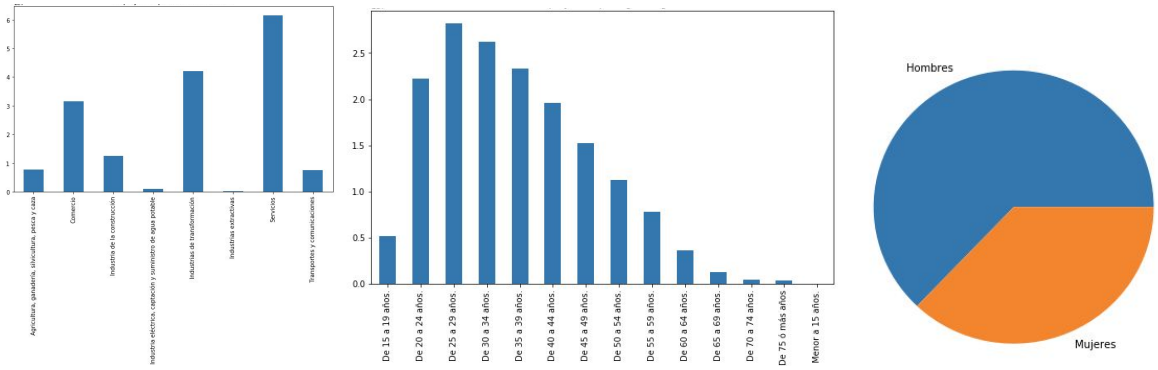
3.

a. Género

i. Representa el género de la población. Puede adquirir los siguientes valores:

1. Hombres
2. Mujeres

Figuras 3,4 y 5: Distribución de variables de segregación respecto a la variable de respuesta



Fuente: Elaboración propia con datos del IIEG

El modelo de predicción consta en utilizar estas variables y el factor temporal para estimar datos a futuro basando en las realizaciones anteriores del nivel de empleo. Por requerimientos administrativos se solicitó la predicción de los siguientes 12 meses a partir del último dato registrado.

Con esta información se generó un *modelo compuesto* por 12 estimadores independientes encargados de realizar la predicción del i-ésimo mes, en dónde *i* representa la posición (orden) del estimador. Tales estimados a su vez son modelos de *machine learning* (*RandomForest*, *Gradient Boosting*) seleccionados por la metodología de *cross validation* mediante un algoritmo de *random search.*[5]

2.2.1.3 Modelo municipal

Dentro de la propuesta de modelo municipal (descrita con detalle en la sección 2.2) se utilizaron 5 modelos (véase 2.1.3.1) de entre los cuales, el código es capaz de discernir entre el que tenga mejor desempeño. Dichos modelos son:

1. SVR (Support Vector Regressor)
2. Random Forest
3. XGBoost (Gradient Boosting con capacidad de multiprocesamiento)
4. Regresión Lasso
5. Regresión Ridge

La segregación utilizada es, como el nombre la sección indica a nivel municipal, donde se prepara un archivo de configuración con parámetros indicados para un *grid search*[6] y el municipio en cuestión, lo cual asegura que para cada municipio se está eligiendo el mejor modelo (de entre los 5 propuestos) de entre muchas más combinaciones (en función de la cantidad de parámetros establecidos).

La propuesta concreta de data para la modelación es una separación por las siguientes características:

[5] Random search: Búsqueda aleatoria en intervalos definidos para una serie de parámetros.
[6] Grid search: Búsqueda exhaustiva de entre una cantidad establecida de parámetros.

1. Rezagos de la serie original definidos por autocorrelación y autocorrelación parcial.
2. Datos estadísticos de la región, como media, desviación estándar y sumas de los empleos en la región (sin considerar el municipio actual) en periodos anteriores.
3. Datos de empleo del municipio en cuestión segmentados por división económica en el dato inmediato anterior.
4. Dato de temporalidad, mes.

Además, dentro de la propuesta municipal, hay 3 vertientes distintas:

1. Modelación mensual t+1
    a. Es un modelo capaz de estimar el mes siguiente teniendo al menos el dato inmediato anterior.
2. Modelación anual t+1
    a. Es un modelo capaz de estimar el año siguiente teniendo al menos el dato inmediato anterior.
    b. Este modelo se hizo con la intención de tener una estimación para todos los componentes del año (meses) mediante una técnica ponderada (véase 2.1.3.1.6).
3. Modelación mensual t+n
    a. Es un modelo capaz de estimar *n* periodos adelante a partir del último punto del tiempo observado.

2.2.2 Plan de trabajo

La metodología de trabajo se puede definir como un ciclo de actividades de desarrollo. Este ciclo consta de los siguientes pasos:

1. **Definición / retroalimentación de requerimientos y funcionalidad por parte de *stakeholders*.**
   Se definen los objetivos y entregables. En siguientes iteraciones se proporciona retroalimentación sobre la metodología adoptada así como de posibles soluciones que la base de la experiencia pudiese incorporar a un algoritmo automatizado.
2. **Planeación de propuesta de solución.**
   Reuniones internas con el equipo de trabajo para definir la mejor forma de abordar los requerimientos. Esto conlleva, realizar validaciones teóricas y análisis de viabilidad.

3. **Generación de prototipo y resultados preliminares.**
   Se genera un prototipo para concretar la propuesta de solución. Se muestran resultados preliminares para validar el error tolerado así como los formatos de las estimaciones.

Habiendo convergido hacia un plan de desarrollo, se procede a generar una solución robusta.

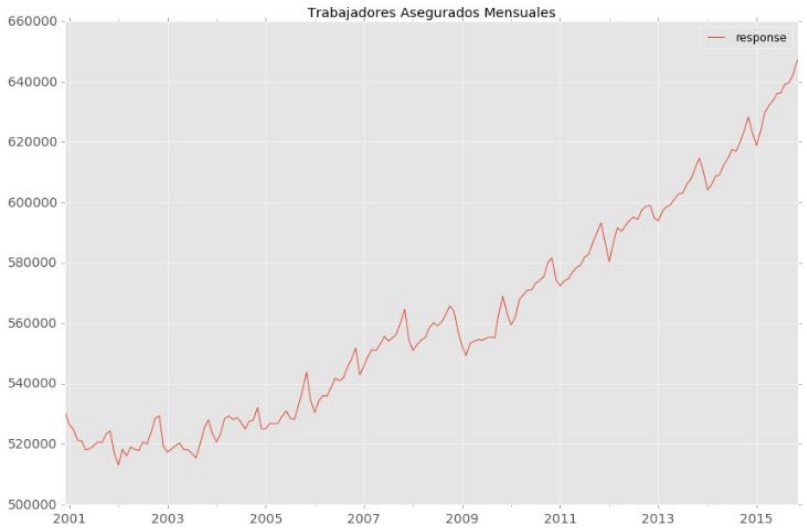
2.2.3 Desarrollo de propuesta de mejora

2.2.3.1 Modelación econométrica

La generación del modelo econométrico se da a nivel municipal con una serie mensual de cada municipio del área metropolitana de guadalajara. Se mostrará el análisis del municipio de Guadalajara para efectos prácticos.

Análisis de gráfica

Figura 6: Empleos formales asegurados en el municipio de guadalajara

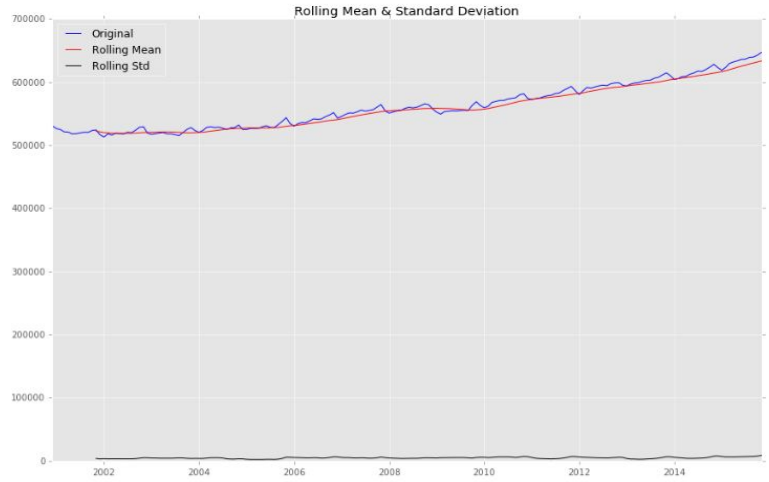


Fuente: Elaboración propia con datos del IIEG

En la gráfica, que muestra los empleados en Guadalajara del 31 de diciembre de 2000 al 31 de se analiza primeramente que existe una tendencia positiva de la misma y tiene ciertos patrones que podrían indicar una estacionalidad en la serie.

Estacionalidad de la serie

Figura 7: Media y desviación estándar a través del tiempo




```
Results of Dickey-Fuller Test:
Test Statistic                  1.807745
p-value                         0.998364
#Lags Used                     12.000000
Number of Observations Used   167.000000
Critical Value (5%)            -2.879008
Critical Value (1%)            -3.470126
Critical Value (10%)           -2.576083
```

Fuente: Elaboración propia con datos del IIEG

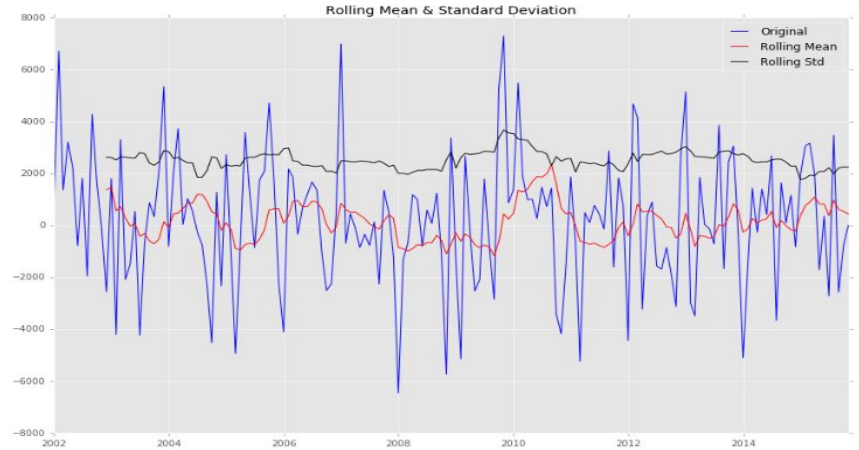
En la figura 7 se observa una gráfica no estacionaria, debido a que la media tiene una tendencia positiva que indica que depende del tiempo, pero se realiza una prueba *Dickey Fuller* Aumentada para comprobarlo, dando como resultado un *p-value* mayor a .05, por lo que no se rechaza la hipótesis nula que establece que la serie no es estacionaria.

Diferencias para convertir la serie en estacionaria

Una vez que se analiza que la serie no es estacionaria, se realizan diferencias en la serie para hacerla estacionaria, y se elige la que tiene la menor desviación estándar. En este caso la que dio una desviación menor fue la primera diferencia aunada con la primera diferencia estacional, lo que nos indica como cierta la teoría de iniciar con un modelo estacional.

Segunda prueba de estacionariedad

Figura 8: Segunda prueba de estacionariedad



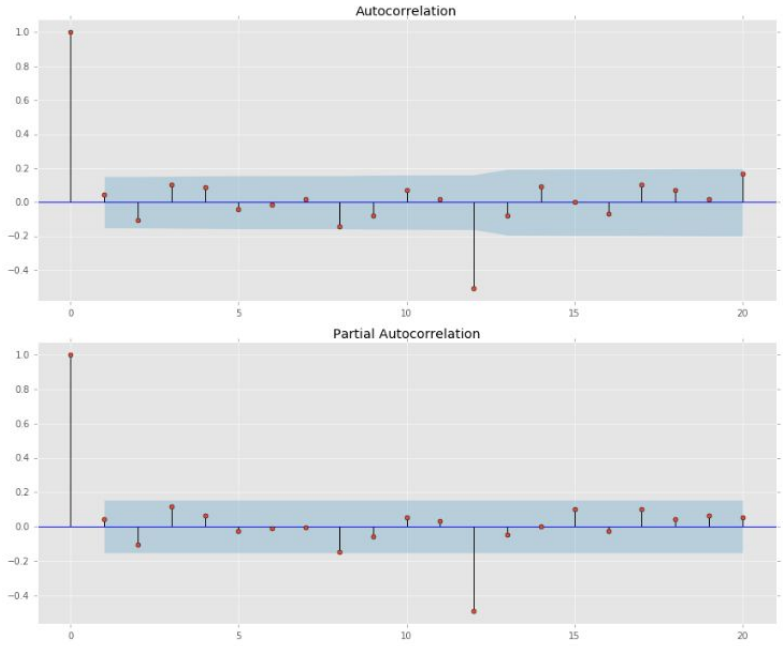

```
Results of Dickey-Fuller Test:
Test Statistic                  -6.516033e+00
p-value                          1.070799e-08
#Lags Used                       1.100000e+01
Number of Observations Used      1.550000e+02
Critical Value (5%)             -2.880374e+00
Critical Value (1%)             -3.473259e+00
Critical Value (10%)            -2.576812e+00
```

Fuente: Elaboración propia con datos del IIEG

La segunda prueba estacionaria se realizó (figura 8) a la serie con una diferencia aunada con diferencia estacional, y esta vez dió un *p-value* menor a .05 por lo que entra en el área de rechazo que establece que la serie es estacionaria, por lo que se realiza con esta serie el ajuste del modelo.

Autocorrelación, autocorrelación parcial y primer modelo

Figura 9: FAC y FAC parcial



Fuente: Elaboración propia con datos del IIEG

Al analizar las gráficas de autocorrelación y autocorrelación parcial se obtienen los parámetros a utilizar con el primer modelo, en este caso se obtuvo un *SARIMA(12,1,12)(1,1,1,12)*.

Tabla 3: Evaluación de significancia en el primer model

Statespace Model Results

Dep. Variable:	response	No. Observations:	180
Model:	SARIMAX(12, 1, 12)x(1, 1, 1, 12)	Log Likelihood	-1263.927
Date:	Tue, 28 Nov 2017	AIC	2581.854
Time:	23:08:12	BIC	2668.064
Sample:	12-01-2000 - 11-01-2015	HQIC	2616.808
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4881	0.170	2.864	0.004	0.154	0.822
ar.L2	-0.3714	0.160	-2.318	0.020	-0.685	-0.057
ar.L3	0.2193	0.130	1.693	0.091	-0.035	0.473
ar.L4	-0.0088	0.162	-0.054	0.957	-0.327	0.309
ar.L5	-0.3932	0.154	-2.549	0.011	-0.696	-0.091
ar.L6	0.3137	0.157	1.994	0.046	0.005	0.622
ar.L7	-0.3831	0.153	-2.497	0.013	-0.684	-0.082
ar.L8	0.0344	0.115	0.299	0.765	-0.191	0.260
ar.L9	0.1057	0.112	0.944	0.345	-0.114	0.325
ar.L10	-0.1551	0.100	-1.557	0.119	-0.350	0.040
ar.L11	0.2323	0.107	2.164	0.030	0.022	0.443
ar.L12	-0.5191	0.096	-5.395	0.000	-0.708	-0.331
ma.L1	-0.4690	0.170	-2.756	0.006	-0.803	-0.136
ma.L2	0.1529	0.137	1.120	0.263	-0.115	0.420
ma.L3	-0.0046	0.124	-0.037	0.970	-0.247	0.238
ma.L4	-0.0194	0.128	-0.151	0.880	-0.270	0.231
ma.L5	0.5222	0.138	3.778	0.000	0.251	0.793
ma.L6	-0.3547	0.135	-2.636	0.008	-0.618	-0.091
ma.L7	0.5333	0.147	3.624	0.000	0.245	0.822
ma.L8	-0.1015	0.091	-1.110	0.267	-0.281	0.078
ma.L9	-0.0143	0.105	-0.137	0.891	-0.219	0.191
ma.L10	0.2136	0.088	2.421	0.015	0.041	0.387
ma.L11	-0.3512	0.112	-3.131	0.002	-0.571	-0.131
ma.L12	0.0762	0.166	0.460	0.645	-0.248	0.401
ar.S.L12	0.2784	0.060	4.647	0.000	0.161	0.396
ma.S.L12	-0.3881	0.051	-7.650	0.000	-0.488	-0.289
sigma2	3.427e+06	1.17e-07	2.94e+13	0.000	3.43e+06	3.43e+06

Ljung-Box (Q):	44.46	Jarque-Bera (JB):	8.66
Prob(Q):	0.29	Prob(JB):	0.01
Heteroskedasticity (H):	1.46	Skew:	-0.21
Prob(H) (two-sided):	0.20	Kurtosis:	4.14

Fuente: Elaboración propia con datos del IIEG

2.2.3.1.1 Reajuste del modelo

Tabla 4: Evaluación de significancia después de un reajuste del modelo.

Statespace Model Results

Dep. Variable:	response	No. Observations:	180
Model:	SARIMAX(6, 1, 5)x(1, 1, 0, 12)	Log Likelihood	-1348.371
Date:	Tue, 28 Nov 2017	AIC	2722.742
Time:	23:32:14	BIC	2764.250
Sample:	12-01-2000 - 11-01-2015	HQIC	2739.572
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	2.1784	0.274	7.955	0.000	1.642	2.715
ar.L2	-2.8928	0.541	-5.346	0.000	-3.953	-1.832
ar.L3	2.6956	0.594	4.536	0.000	1.531	3.860
ar.L4	-1.8942	0.433	-4.373	0.000	-2.743	-1.045
ar.L5	0.9530	0.207	4.614	0.000	0.548	1.358
ar.L6	-0.1593	0.056	-2.864	0.004	-0.268	-0.050
ma.L1	-2.1383	0.263	-8.141	0.000	-2.653	-1.623
ma.L2	2.7275	0.570	4.786	0.000	1.611	3.844
ma.L3	-2.4044	0.637	-3.777	0.000	-3.652	-1.157
ma.L4	1.5875	0.499	3.179	0.001	0.609	2.566
ma.L5	-0.7292	0.235	-3.097	0.002	-1.191	-0.268
ar.S.L12	-0.5473	0.076	-7.238	0.000	-0.695	-0.399
sigma2	4.615e+06	1.89e-07	2.44e+13	0.000	4.61e+06	4.61e+06

Ljung-Box (Q):	54.85	Jarque-Bera (JB):	0.31
Prob(Q):	0.06	Prob(JB):	0.85
Heteroskedasticity (H):	0.88	Skew:	-0.05
Prob(H) (two-sided):	0.65	Kurtosis:	3.20

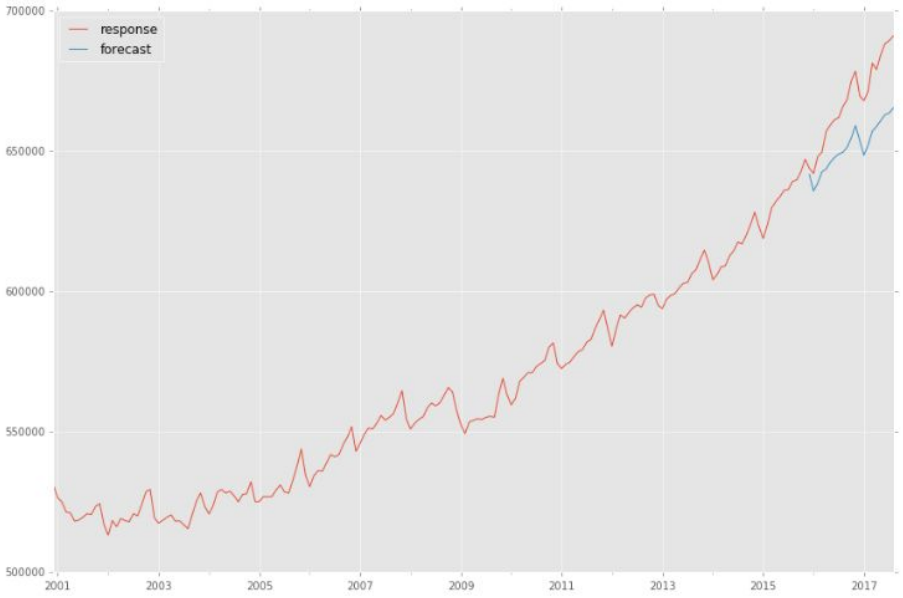
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.12e+30. Standard e:

Fuente: Elaboración propia con datos del IIEG

Se eliminaron los parámetros que no eran significativos al primer modelo, para generar un modelo que tuviera una mejor adaptación, dando un modelo *SARIMA(6,1,5)(1,1,0,12).*

Predicción del modelo

Figura 10: Estimación del modelo

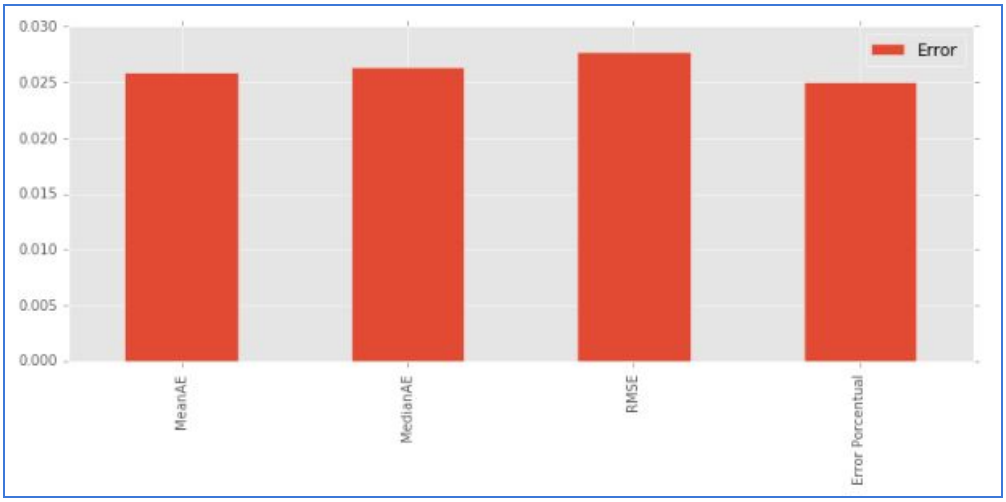


Fuente: Elaboración propia con datos del IIEG

Se realizó una predicción con el modelo de enero 2016 a agosto 2017 para probar la precisión del modelo, y posteriormente se analizaron los errores para analizar el poder de predicción del modelo.

Métricas de error del modelo

Figura 11: Métricas del error en el modelo



Fuente: Elaboración propia con datos del IIEG

En el caso de Guadalajara se obtuvieron errores con un porcentaje promedio de 2.5%, en el caso del error porcentual se podría decir que se esperá un 2.5% de variación del resultado de la predicción en relación con los datos reales de empleo.

2.2.3.2 Modelación *machine learning*

2.2.3.2.1 Modelación estatal

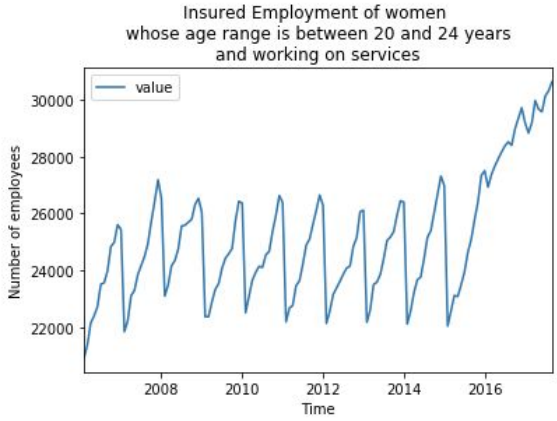
El modelo de estimación del nivel de empleo estatal se basa en el nivel de segregación previamente descrito. Cada uno de los 12 estimadores que conforman el modelo compuesto sigue una metodología de determinación de rezagos óptimos explicada a continuación:

Rezagos óptimos

Para crear el dataset de modelación para 1 estimador se debe definir el número de rezagos a utilizar. El rezago mínimo aceptable es definido en base al “orden” del estimador (el número de periodos temporales para estimar a futuro). Dado estos rezagos, para estimar los rezagos significativos se debe analizar cada serie de tiempo posible dado las realizaciones de las variables de segregación. Por ejemplo, una serie de tiempo podría corresponder a las mujeres (género) entre 20 y 24 años (rango de edad) trabajando en servicios (división económica).

La siguiente gráfica muestra los resultados de este nivel de segregación:

Figura 12: Nivel de empleo asegurado para cierto nivel de segregación



Fuente: Elaboración propia con datos del IIEG

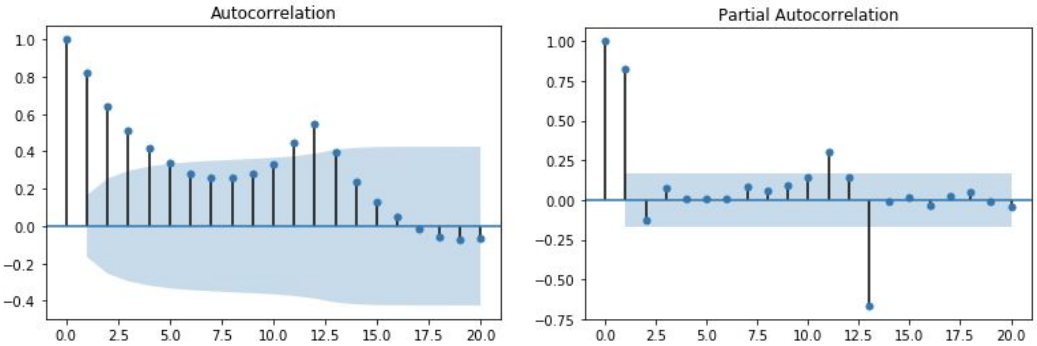
Tabla 5: Ejemplo de vista de datos para cierto nivel de segregación

	economic_division	gender	age_range	value	year	month	time
174	Servicios	Mujeres	De 20 a 24 años.	20908.0	2006.0	1	2006.083333
388	Servicios	Mujeres	De 20 a 24 años.	21362.0	2006.0	2	2006.166667
602	Servicios	Mujeres	De 20 a 24 años.	22139.0	2006.0	3	2006.250000
817	Servicios	Mujeres	De 20 a 24 años.	22397.0	2006.0	4	2006.333333
1033	Servicios	Mujeres	De 20 a 24 años.	22712.0	2006.0	5	2006.416667

Fuente: Elaboración propia con datos del IIEG

Utilizando técnicas econométricas (FAC y FACP) se determina los rezagos significativos de esta serie. Suponiendo que estamos generando los rezagos para un estimador de orden 1 (donde el rezago mínimo es t-1) tenemos:

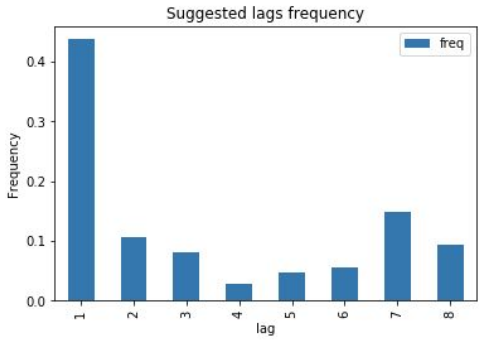
Figura 13: Funciones FAC y FACP



Fuente: Elaboración propia con datos del IIEG

Se realiza este mismo análisis para todas las series de tiempo generadas por las combinaciones posibles de las variables de segregación. Se procede a visualizar el resultado como una gráfica de frecuencia y se seleccionaron aquellos rezagos que contribuyeron en más del 5% de todas las series.

Figura 14: Gráfica de frecuencia de uso de rezagos para cada serie generada por el nivel de segregación (para un estimador de orden 1)



Fuente: Elaboración propia con datos del IIEG

Como resultado se agregan los rezagos correspondientes a los datos. Este nuevo dataset es utilizado para el proceso de entrenamiento del estimador 1. El mismo procedimiento se repite para los estimadores faltantes.

Tabla 6: Dataset base para estimador 1

Available data: (1720, 11)

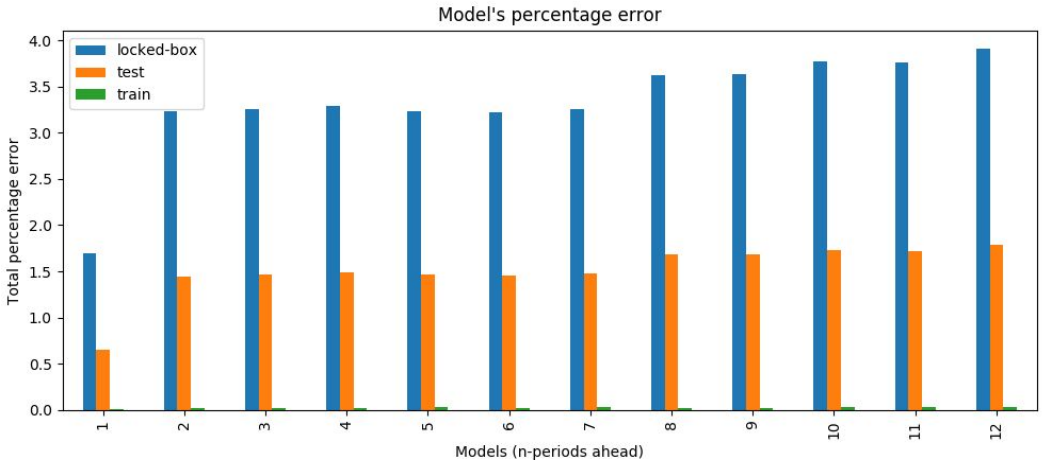
	economic_division	gender	age_range	value	year	month	t-1	t-2	t-3	t-6	t-7
0	Agricultura, ganadería, silvicultura, pesca y ...	Hombres	De 15 a 19 años.	3175.0	2017.0	1	3079.0	2995.0	2882.0	2502.0	2577.0
1	Industrias extractivas	Hombres	De 20 a 24 años.	283.0	2017.0	1	297.0	341.0	349.0	340.0	343.0
2	Industrias extractivas	Hombres	De 25 a 29 años.	458.0	2017.0	1	488.0	537.0	536.0	513.0	499.0
3	Industrias extractivas	Hombres	De 30 a 34 años.	406.0	2017.0	1	406.0	453.0	464.0	466.0	470.0
4	Industrias extractivas	Hombres	De 35 a 39 años.	389.0	2017.0	1	409.0	429.0	427.0	429.0	424.0

Fuente: Elaboración propia con datos del IIEG

Modelo compuesto

El modelo de *machine learning* usado dentro de cada estimador se genera usando la metodología de random search para los algoritmos de Random Forest y Boosted Trees. La siguiente figura muestra el error porcentual agrupado mensualmente generado por tales modelos.

Figura 15: Error porcentual de *train, test, locked-box* para modelo estatal compuesto

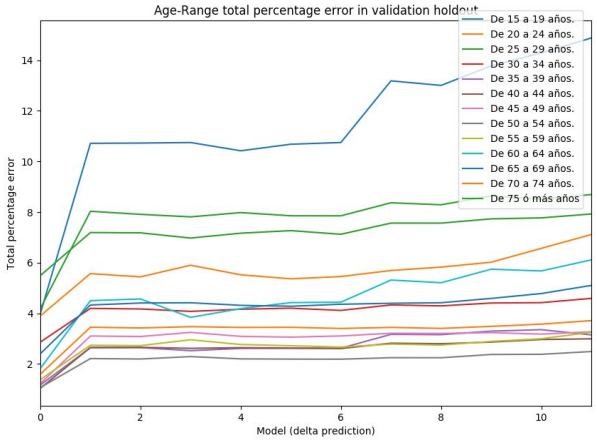


Fuente: Elaboración propia con datos del IIEG

Se puede observar que el modelo presenta un error bastante bajo en los datos de entrenamiento pero aumenta en los datos de validación (*locked-box*). Así mismo, el error en este conjunto aumenta según el orden del modelo. Estos resultados son congruentes debido a que es de esperarse que el poder predictivo de un modelo dado disminuya cuando se solicita estimar a cada vez más periodos temporales por anticipado. El error máximo generado es menor del 4%.

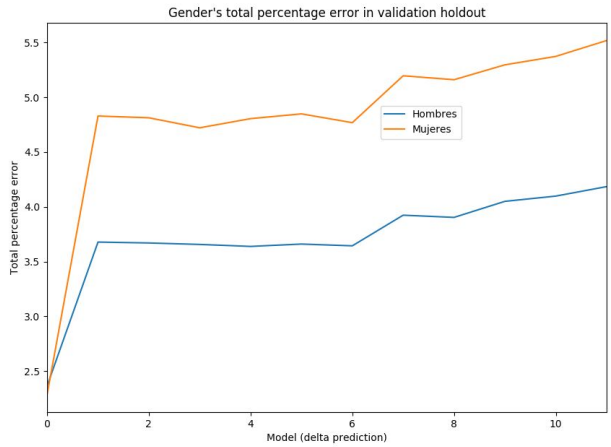
Debido a que estamos manejando datos con un nivel de segregación relativamente alto, es posible mostrar la distribución del error en distintas agrupaciones.

Figura 16: Error porcentual agrupadas por rango de edad



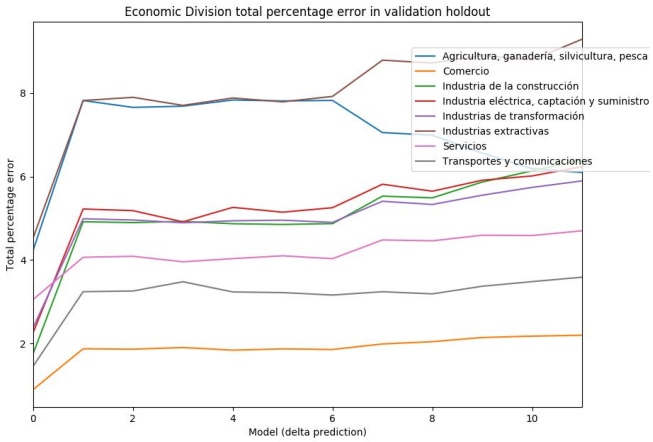
Fuente: Elaboración propia con datos del IIEG

Figura 17: Error porcentual agrupado por género



Fuente: Elaboración propia con datos del IIEG

Figura 18: Error porcentual agrupado por división económica



Fuente: Elaboración propia con datos del IIEG

Los resultados del error son congruentes con los resultados de la agrupación general. Se puede notar como aumenta el error cuando incrementa el orden del modelo. Así mismo, esta visualización muestra la distribución del error dentro de las diferentes categorías de una variable. Por ejemplo, el error que se obtiene al predecir el género *Mujer* es mayor que el de los *Hombres*. Este tipo de precisión generalmente es definido por la calidad de los datos en una categoría. Es común encontrar que aquellas categorías con bajo desempeño tienen pocos datos asignados.

2.2.3.2.2 Modelación municipal

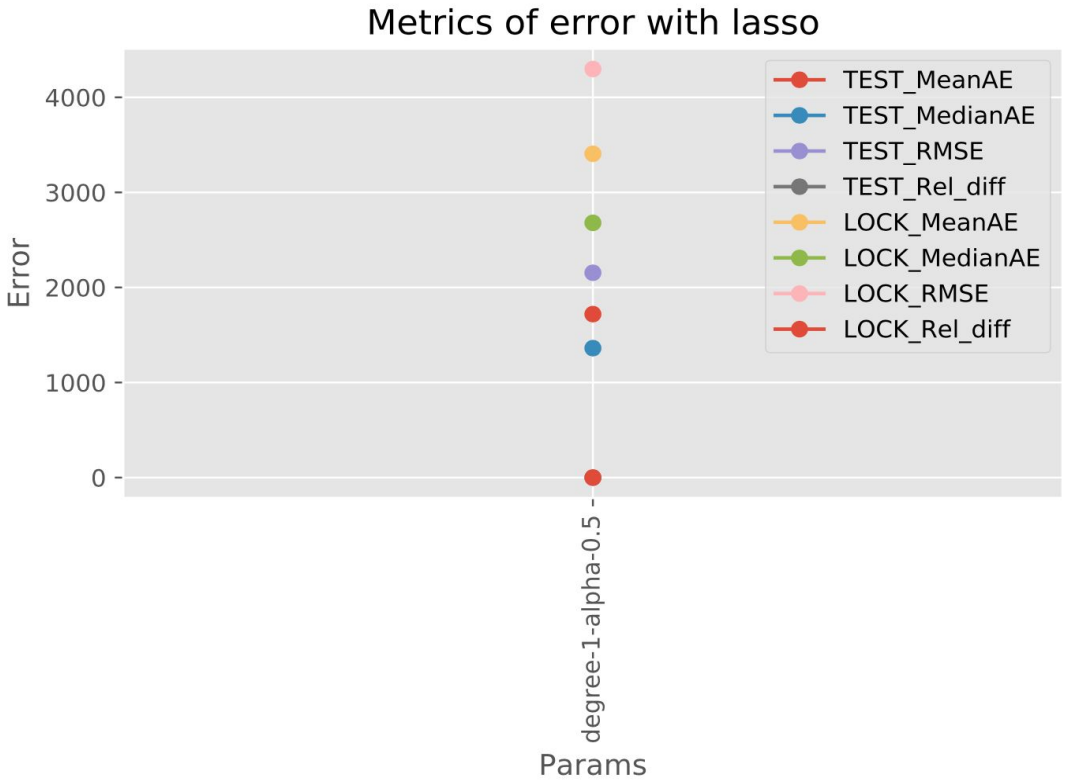
La evaluación de estos modelos se basa en la implementación propuesta en 2.2.1.3 y las métricas de error descritas en 2.1.1.

Como ha sido mencionado en puntos anteriores, se realiza una búsqueda exhaustiva de hiper parámetros obteniendo como resultado la figura 19. Con los datos obtenidos en el conjunto de prueba (*test)* se eligen los hiper parámetros adecuados para obtener una gráfica como la figura 20.

Figura 19: Métricas de error en el modelo Lasso para el municipio de Guadalajara

Fuente: Elaboración propia con datos del IIEG

Figura 20: Métricas de error en el mejor modelo Lasso para el municipio de Guadalajara

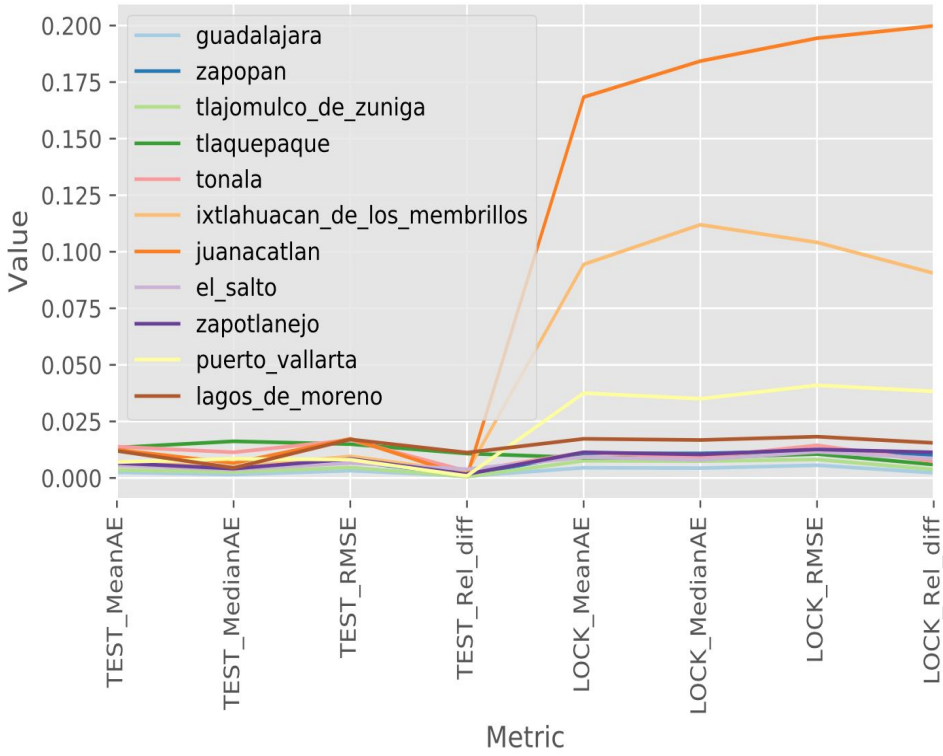


Fuente: Elaboración propia con datos del IIEG

En el caso del municipio de Guadalajara, se define al mejor modelo *Lasso* como el de una regresión de grado 1 con un grado de penalización en *L1* de 0.5. Este proceso de selección de modelos es acumulado durante el código a través de todas las operaciones posibles para, de manera final, discernir entre el mejor modelo de entre los 5 (que a su vez es el mejor de entre una serie de combinación de hiper parámetros).

Con la repetición de este proceso en todo el área metropolitana, se llega a un conjunto de modelos (no necesariamente iguales entre sí) que son el mejor encontrado para el AMG llegando a un error preliminar como el siguiente:

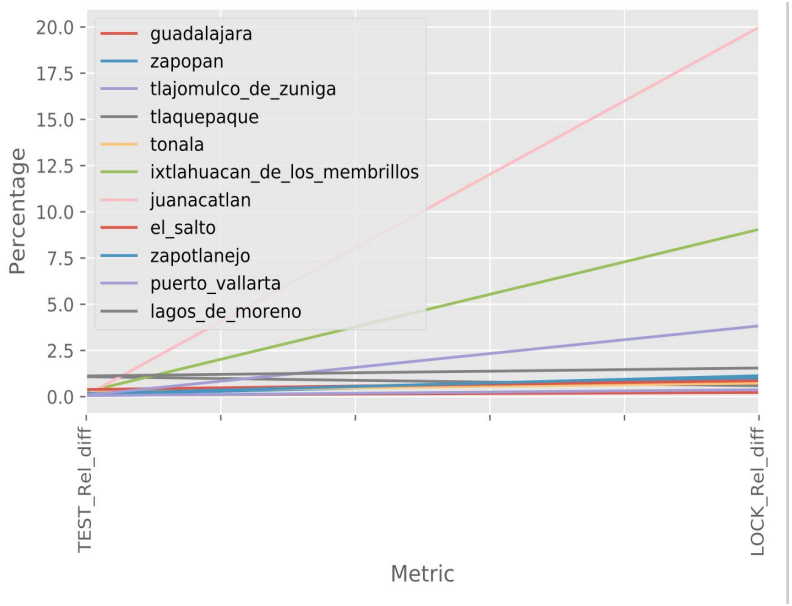
Figura 21: Métricas de error en el mejor modelo de cada municipio para todo el AMG



Fuente: Elaboración propia con datos del IIEG

Por su parte, el error porcentual toma la siguiente forma:

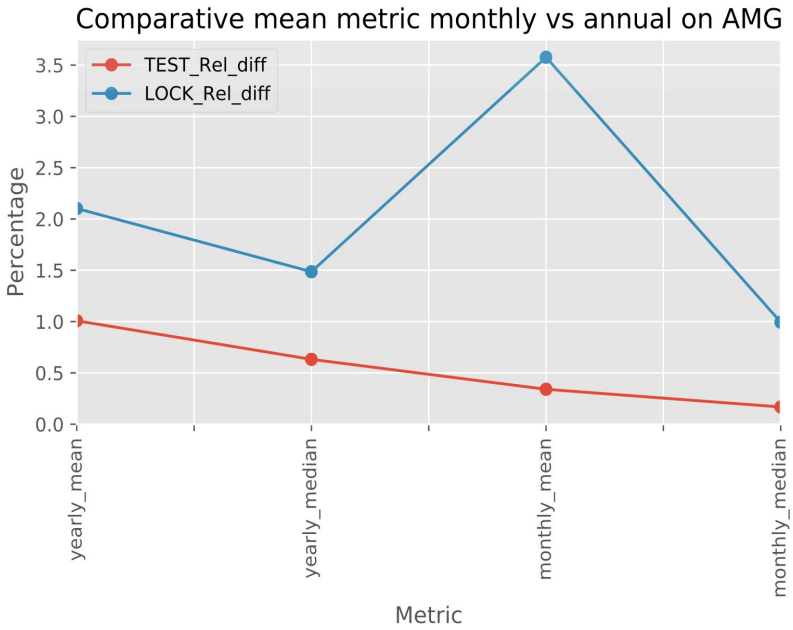
Figura 22:: Métricas de error porcentual en el mejor modelo de cada municipio para todo el AMG



Fuente: Elaboración propia con datos del IIEG

En base a esta decisión y al hallazgo de tener la certeza de un comportamiento con un error razonable en distintas métricas y conjuntos, se llega a la comparación final de modelos, recordando que en el modelo municipal existen tres posibles combinación de modelos.

Figura 23: Métricas de error porcentual en el mejor modelo de cada municipio para todo el AMG

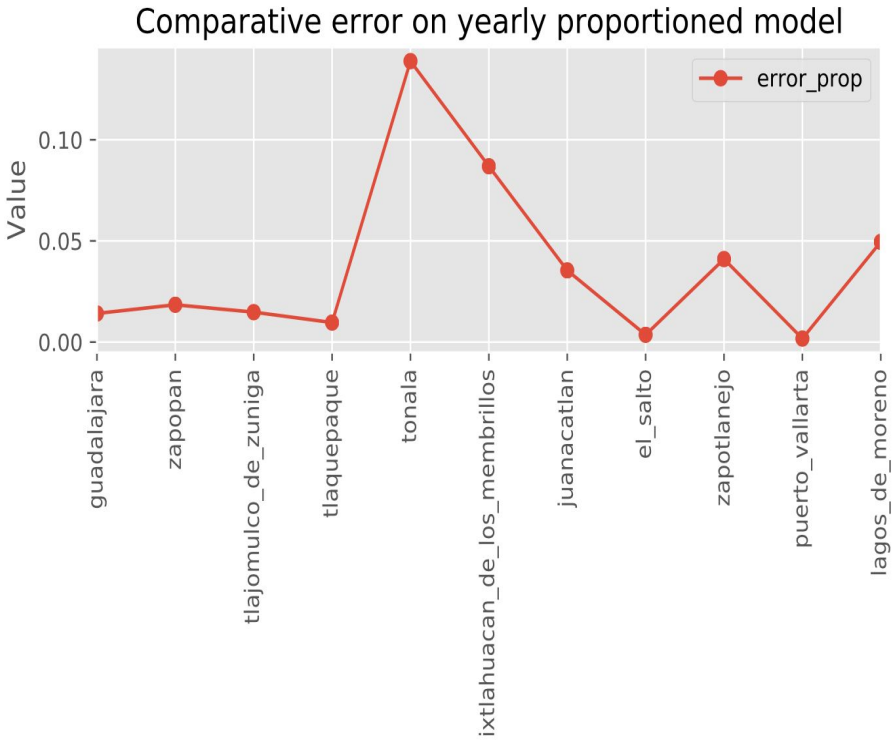


Fuente: Elaboración propia con datos del IIEG

Es posible observar que los *outliers* representados en la figura anterior elevan el promedio del modelo mensual, pero que aun con todo eso, el modelo mensual *t+1* tiene un error de entre 1% (mediano) y 3.5% (promedio). El caso del modelo anual es menormente afectado por este tipo de cuestiones, teniendo errores máximos apenas superiores al 2% en promedio.

De usarse la técnica de ponderación para estimar todo ciertos meses (todos los componentes de un año), se podría esperar un error mediano de 1.84% y un error promedio de 3.76%, errores bastante aceptables comparados a la métrica de 10% mencionado por el IIEG.

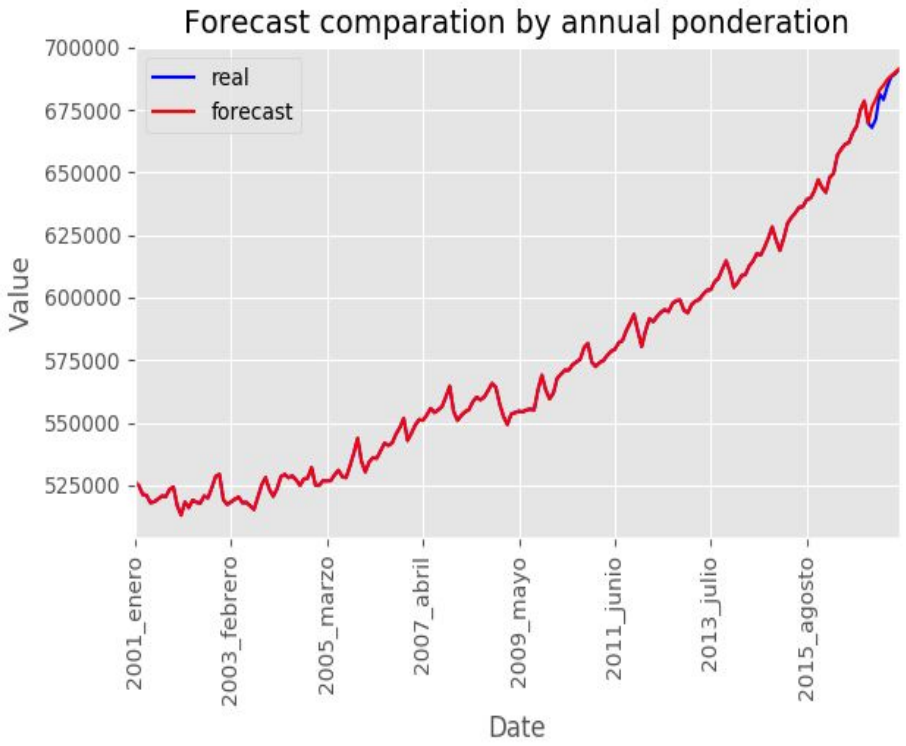
Figura 24: Error comparativo porcentual en la técnica de ponderación por cada municipio para todo el AMG



Fuente: Elaboración propia con datos del IIEG

Un ejemplo de la figura anterior es poner a prueba el modelo anual con su técnica de ponderaciones en algún municipio que, por motivos de simplicidad y espacio, será Guadalajara:

Figura 25: Estimación en 2017 con el modelo anual y su técnica de ponderaciones en Guadalajara



Fuente: Elaboración propia con datos del IIEG

En este punto, es posible hacer la comparación entre esta técnica y la clásica (vista econométrica):

Tabla 7: Comparativa de errores entre modelos econométricos y modelos machine learning en el AMG

	Error promedio	**Error mediano**
Referencia	10%	10%
Econométrico[7]	5.02%	4.52%
Municipal	3.76%	1.84%

Fuente: Elaboración propia con datos del IIEG

[7] Falta Juanacatlán por la imposibilidad econométrica de estimación.

Teniendo entonces la evaluación concreta y los resultados esperados en este tipo de modelación, se muestran algunos resultados (véase PDF anexo para los completos) en la sección 3.2.1.

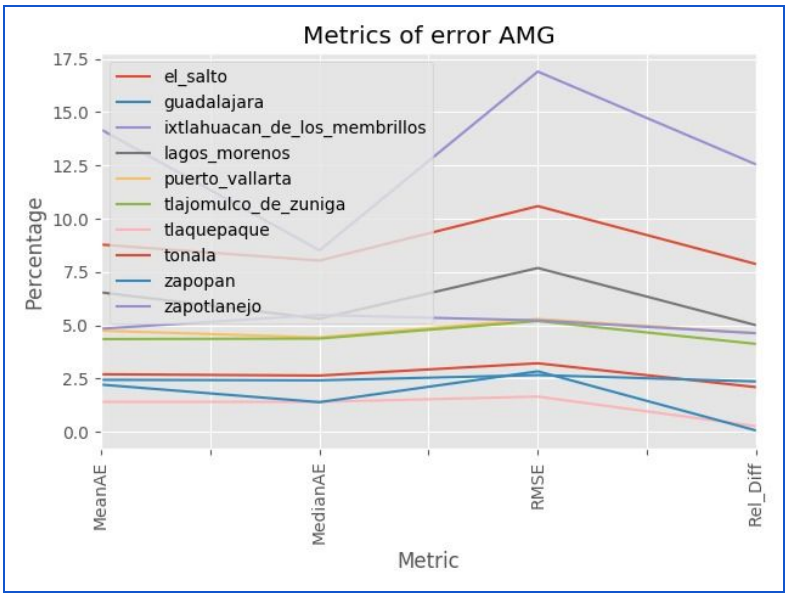
3. Resultados del trabajo profesional

En este punto del trabajo han sido mostradas ya distintas métricas de error en los cuales es posible evaluar el desempeño esperado en conjuntos OOS[8] de los distintos modelos propuestos por los autores, por lo que se muestran a continuación distintos pronósticos en las series. Como nota importante, para poder visualizar todos los resultados esperados para 2018, véase anexos.

3.1 Modelación econométrica

A nivel área metropolitana se obtuvieron los siguientes errores con modelos econométricos ajustados a cada municipio.

Figura 26: Métricas de error en el área metropolitana de Guadalajara con modelación econométrica



Fuente: Elaboración propia con datos del IIEG

Observando el error a nivel área metropolitana, podemos observar que en algunos casos el error que se podría obtener con un modelo econométrico ajustado a cada serie puede llegar a ser muy grande dependiendo de las características de la serie, en este caso, Ixtlahuacan de los membrillos obtuvo un error cuadrático medio de 17% aproximadamente.

[8] Out of sample, fuera de la muestra.

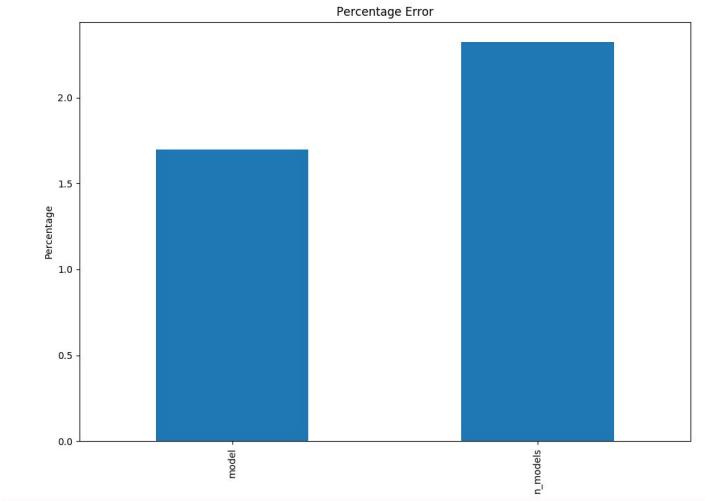
3.2 Modelación *machine learning*

3.2.1 Modelación estatal

El modelo compuesto a nivel estatal es capaz de generar la predicción del nivel de empleo para los siguientes 12 meses de acuerdo al nivel de segregación definido por las variables de **División Económica**, **Rango de Edad** y **Género**.

El error en "producción" porcentual de una estimación mensual (total) para 1 año se puede calcular con el error generado en el conjunto *locked-box.* Obtenemos los siguientes resultados:

Figura 27: Error porcentual para estimación de 12 meses con
modelo de actualización mensual (model) y modelo compuesto (n_model)

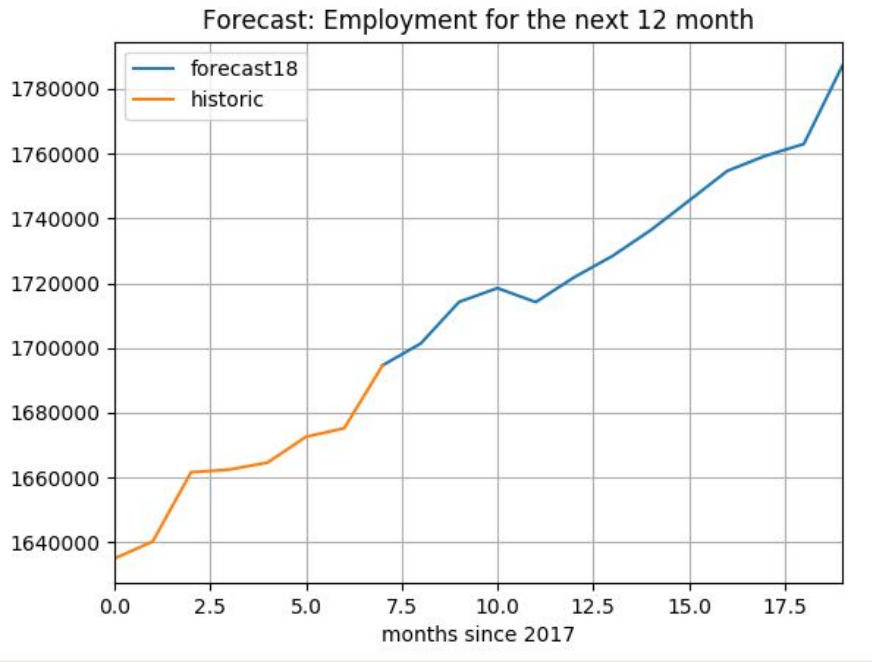


Fuente: Elaboración propia con datos del IIEG

Por lo tanto, el error de las estimaciones mensuales para el periodo de 12 meses es 2.34%. Esto quiere decir que en promedio, el pronóstico de empleo asegurado para cada mes puede diferir 2.34% del dato real.

Bajo este contexto, se presenta la agrupación mensual de las estimaciones para 12 meses a futuro desde el último dato disponible (agosto 2017).

Figura 28: Pronóstico para 12 meses a futuro a nivel estatal



Fuente: Elaboración propia con datos del IIEG

Se puede observar que el pronóstico sigue la tendencia alcista que la serie mostraba previamente. De forma adicional, la ventaja de utilizar un modelo con un alto nivel de segregación es que nos permite “agrupar” el pronóśtico de forma arbitraria según las variables de interés.

Cada uno de los 12 estimadores genera un set de datos como pronóstico en donde cada “renglón” o “fila” representa el pronóstico para tal nivel de segregación. A continuación se muestran algunos resultados de pronóstico del primer estimador:

Tabla 8: Resultados del pronóstico del estimador 1

	age_range	economic_division	gender	month	time	value	year	t-10	t-11	t-12	t-13	t-1	forecast
0	De 15 a 19 años.	Agricultura, ganadería, silvicultura, pesca y ...	Hombres	9	2017.666667	0	2017.0	2995.0	2882.0	2610.0	2479.0	2809.0	2993.847466
1	De 15 a 19 años.	Agricultura, ganadería, silvicultura, pesca y ...	Mujeres	9	2017.666667	0	2017.0	1424.0	1422.0	1278.0	1247.0	1436.0	1450.453192
2	Menor a 15 años.	Agricultura, ganadería, silvicultura, pesca y ...	Mujeres	9	2017.666667	0	2017.0	4.0	3.0	3.0	2.0	4.0	2.565112
3	Menor a 15 años.	Agricultura, ganadería, silvicultura, pesca y ...	Hombres	9	2017.666667	0	2017.0	4.0	3.0	2.0	2.0	2.0	1.812899
4	De 75 ó más años	Agricultura, ganadería, silvicultura, pesca y ...	Mujeres	9	2017.666667	0	2017.0	220.0	223.0	216.0	202.0	182.0	195.782282
5	De 75 ó más años	Agricultura, ganadería, silvicultura, pesca y ...	Hombres	9	2017.666667	0	2017.0	469.0	459.0	452.0	442.0	467.0	471.009576
6	De 65 a 69 años.	Agricultura, ganadería, silvicultura, pesca y ...	Mujeres	9	2017.666667	0	2017.0	320.0	316.0	310.0	299.0	291.0	294.574723
7	De 65 a 69 años.	Agricultura, ganadería, silvicultura, pesca y ...	Hombres	9	2017.666667	0	2017.0	945.0	922.0	888.0	863.0	901.0	854.840050

Fuente: Elaboración propia con datos del IIEG

La columna de *value* es cero debido a que no se tiene el valor real en ese tiempo. La columna *forecast* representa la estimación para dentro de 1 mes dado el valor en el Rango de Edad, División Económica y Género.

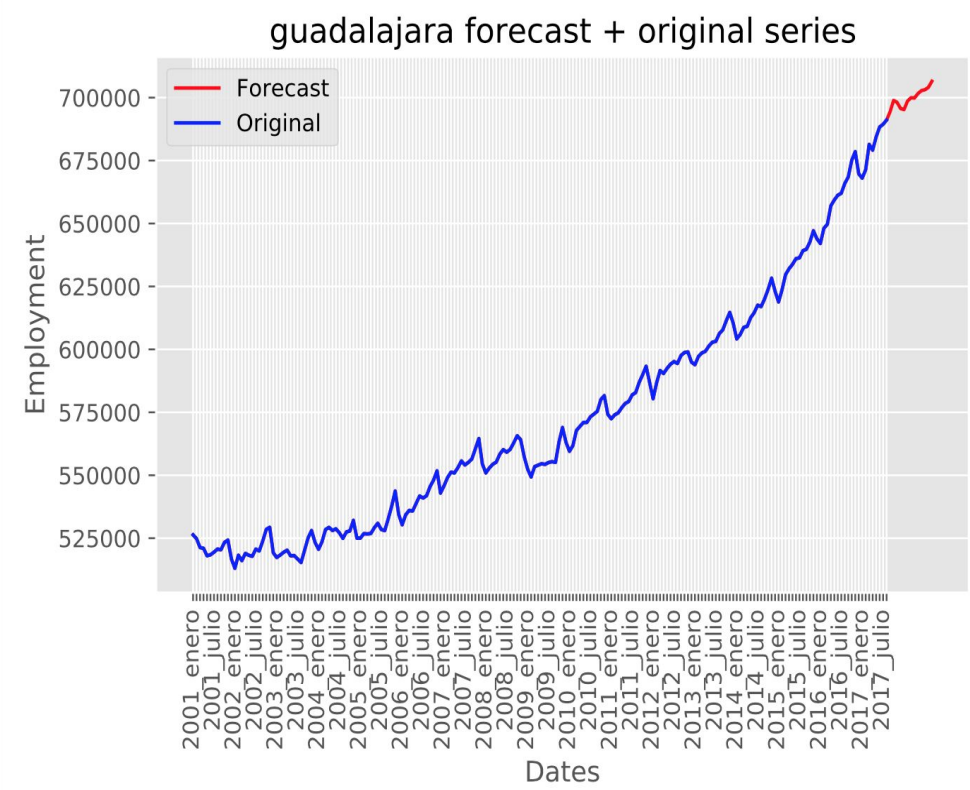
De esta forma, es posible extraer estimaciones para una categoría (o conjunto de categorías) de especial interés a nivel estado.

3.2.2 Modelación municipal

El caso del área metropolitana es complejo y de muchas series particulares que resultaría difícil analizar en una sola figura y sobresaturado adjuntar distintas figuras con estas predicciones para toda el área, por lo que por motivos de espacio se muestran los pronósticos de la serie de Guadalajara.

La siguiente figura muestra el pronóstico realizado en la ya mencionada capital del Estado:

Figura 29: Pronóstico en la ciudad de Guadalajara con un modelo t+n.



Fuente: Elaboración propia con datos del IIEG

Más allá de los pronósticos que puedan mostrarse para las series municipales en general, el producto obtenido es un código que puede discernir entre la metodología de modelación adecuada (los distintos modelos con las distintas variedades de hiper parámetros) y el distinto criterio original (mensual *t+1*, anual *t+1* con su debida proporción o *t+n*), arrojando en material gráfico como el aquí expuesto y en números puntuales (tablas, estimaciones) las predicciones para el periodo y región de interés.

4. Reflexiones del alumno o alumnos sobre sus aprendizajes, las implicaciones éticas y los aportes sociales del proyecto

4.1 Aprendizajes profesionales

Durante el desarrollo de un proyecto como el que, en conjunto con el Instituto acabamos de realizar, se ponen en práctica muchos de los valores propuestos por el ITESO en su plan de estudios que dan ese sello profesional a todos los egresados por la universidad. Además de eso, se llevan al campo laboral conocimientos básicos pero conectados por un mismo producto, conocimientos como:

1. Probabilidad y estadística aplicadas.
2. Programación aplicada a modelos probabilísticos de corte y proyección de tipo regresor.
3. Cálculo básico para entender los procesos de optimización en las funciones de coste.

El poder tener esa influencia sobre la estimación de un dato tan importante para la decisión del manejo y creación de políticas públicas relevantes en una región mediante aprendizajes *cotidianos* de ingeniería financiera es lo que nos trajo a cada uno de nosotros a este punto, el apoyo y la aplicación de una ingeniería para la solución de un proceso complejo en una zona determinada.

4.2 Aprendizajes sociales

Cada uno de nosotros autores como profesionistas, vemos la culminación de nuestra carrera como una aplicación laboral sobre los aprendizajes obtenidos que más hayamos disfrutado durante los estudios, sin embargo, en ningún momento -al menos a nosotros en particular- nos cruzó por la mente la posibilidad de que en una región desarrollada como la que habitamos -Guadalajara- pudiese ocurrir lo impensable: No hay empleo.

Ese tipo de problemas se pueden evitar -aunque no asegurar- con, entre muchos factores involucrados, la creación de políticas públicas adecuadas que permitan a la entidad desarrollarse o especializarse y por ende tener empleo y consecuentemente crecimiento.

Ese granito de arena que estamos cada uno de nosotros aportando con el desarrollo de estos modelos permite la identificación de problemas de temporalidad o de un comportamiento errático y a su vez, permitir que las autoridades correspondientes tomen análisis correspondientes en la atracción de inversión, en la creación de empleos de cierto tipo o cierto sector o inclusive, en la constante mejora por la competencia contra entidades que no tengan herramientas de este tipo.

4.3 Aprendizajes éticos

La ética en el proyecto es crucial porque, a pesar de que no se está trabajando con información estrictamente confidencial, el mal uso o una aplicación irresponsable de los datos podrían llevar a conclusiones al fin, pero conclusiones erróneas que podrían empezar con un diagnóstico equivocado y un error constantemente creciente. Por esto mismo las verificaciones de correcta programación y estadísticamente significativos son algo que nos encargamos de realizar con alto grado de formalidad.

4.4 Aprendizajes en lo personal

Sin lugar a la menor duda, este proyecto se encargó de proyectar no solamente los empleos, si no además de un sentido de constante revisión de un trabajo conjunto que sea significativo, reproducible pero sobre todo ajustable para condiciones cambiantes.

Es por eso que tomamos las decisiones expuestas anteriormente tienen un significado en cuanto a datos se refiere, pero también en la vida de cada uno, pues implicó un proceso constante de investigación y aplicación de metodologías de vanguardia que fueran las idóneas para el ajuste de las series en cuestión, fuese la segregación que fuese la que se estuviera analizando.

5. Conclusiones

Durante el desarrollo de este proyecto de aplicación profesional pudimos aprender la importancia del manejo correcto de los datos y el valor e impacto que tienen en la toma de decisiones. Implementar modelos de “machine learning” en el “estado del arte” y desarrollar una solución general con nivel de abstracción suficiente para presentarlo y usar los resultados

a nivel administrativo generó un reto integrador que requirió un alto desempeño en distintas áreas fundamentales de nuestra carrera.

Consideramos que en general se cumplieron los objetivos y requerimientos propuestos por el Instituto de Información Estadística y Geográfica ya que se logró una implementación robusta de varios modelos predictivos para la variable de empleos asegurados con un nivel de error menor que los modelos que se utilizaban anteriormente.

6. Bibliografía

- De la Fuente Fernández, S. (2017). Series Temporales: Modelo Arima. Obtenido de Portal Estadística Aplicada. Consultoría Estadística-Econometría.: http://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf
- González Casimiro, M. P. (2017). *Análisis de series temporales: Modelos Arima.* Obtenido de Universidad del país Vasco: https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf;jsessionid=2A515AF2E29E6DC94BE9B78786C8BC99?sequence=1
- Mentz, R. (1988). Estimación de los Modelos Autorregresivos y de Promedios Móviles. *Estadística Española*, 87-106. Obtenido de http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/SeriesTemporales/Tema4SeriesEstud.pdf
- Scikit Learn (2017). Ensemble methods. Retrieved from http://scikit-learn.org/stable/modules/ensemble.html#forest
- Scikit Learn (2017). Linear models. Retrieved from http://scikit-learn.org/stable/modules/linear_model.html#ridge-regression
- Sickit Learn. (2017) Support Vector Regressor. Retrieved from http://scikit-learn.org/stable/modules/svm.html#svm-regression
- Scikit Learn. (2017). Linear models. Retrieved from http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html
- Tsay, R. S. (2005). *Analysis of Financial Time Series.* Canadá: John Wiley & Sons, Inc. Publication.
- Utah State University A. C. (2013, October 3). Trees and Random Forests. Retrieved from http://www.math.usu.edu/adele/RandomForests/UofU2013.pdf