Tráfico de Internet: Una Aplicación de Datos Georreferenciados y Series de Tiempo

Florencia De Arca*[1], Carlos Germán Carreño Romano*†[2], Claudio Righetti*[3] and Gabriel Carro*[4]
*Telecom Argentina S.A.,
Alicia Moreau de Justo 50, C1107AAB, Buenos Aires, Argentina
†Universidad de Buenos Aires, Facultad de Ingeniería,
Departamento de Física, GLOmAe, Buenos Aires C1063ACV, Argentina
[1] fdearca@teco.com.ar
[2] caromano@teco.com.ar
[3] crighetti@teco.com.ar
[4] gcarro@teco.com.ar

***Resumen*—En este trabajo se presentan análisis y resultados de datos de uso de Internet georreferenciados para Argentina en el período 2017-2019 tomados de la red de Telecom Argentina S.A. Algunas variables de estudio que nos permiten obtener información acerca de comportamiento de usuarios son el tráfico de datos en la hora pico, su evolución en el tiempo, el ancho de banda contratado, entre otros. Parte de esta información suele ser presentada en reportes de manera periódica dentro de las compañías que brindan acceso a Internet. Sin embargo, se cuenta con demasiada información como para poder plasmarla en un único reporte de manera sintética. Los resultados de este trabajo muestran cómo con arquitecturas de código abierto se desarrolla una aplicación interactiva que presenta toda la información disponible con gran detalle de la capilaridad de la red. Por último, se presentan métodos de predicción como una aplicación a las series de tiempo de tráfico que permiten dar idea de la red del futuro.**

***Index Terms*—Georeferencing, time series, Internet traffic, data processing, forecast.**

Abstract

This paper presents analysis and results of georeferenced Internet usage data for Argentina in the period 2017-2019 taken from the Telecom Argentina S.A. network. Some variables which allow us to obtain information about user behavior are data traffic at peak time, its evolution over time, the contracted bandwidth, among others. This information is usually presented in reports on a regular basis within companies that provide Internet access. However, there is too much information which cannot be presented in a single report in a synthetic way. The results of this work show how, with open source architectures, an interactive application is developed that presents all the available information with great detail of the capillarity of the network. Finally, prediction methods are presented as an application to the traffic time series that give an idea of the network of the future.

I. Introducción

La industria de las telecomunicaciones y servicios digitales está experimentando una transformación. Los servicios de conectividad han posibilitado el tendido de grandes redes de acceso y transporte de datos, tanto para telefonía fija o móvil como para la distribución de contenidos de televisión o radio, entre otros servicios. La convergencia de los servicios sobre la misma red va dando lugar a la aparición de nuevas arquitecturas y funcionalidades unificadas. Las tendencias de NFV/SDN [1] plantean abstracciones que llevan al dominio del software funciones que muchas veces son de hardware específico. En esa línea, la automatización de procesos de despliegue de infraestructura, el análisis masivo de datos y el desarrollo de herramientas adecuadas para propósitos distintos como pueden ser el análisis de fallas, la detección de anomalías, la información en tiempo real, el modelado de indicadores, entre otros, aborda en la actualidad enfoques de aprendizaje automático o de inteligencia artificial [2]. Los datos entonces pasan a ser la materia prima del futuro de los servicios de conectividad, entendiendo que la infraestructura física de red ya ha sido mayormente desplegada.

II. Metodología

Al estudiar los datos aparecen múltiples dimensiones: variables en función del tiempo, en función del espacio, en función de alguna variable de control como puede ser un identificador, un producto contratado, un modelo de equipamiento instalado, o en función de un historial de fallas físicas, lógicas, programadas, etc. por mencionar sólo algunas. Algunas fuentes de datos que representan variables de negocio suelen ser más o menos estáticas, pero otras referidas al comportamiento de usuario son bien dinámicas y hace falta encontrar los puntos de medición adecuados para lograr precisión en la información. Estos orígenes de datos pueden ser, por ejemplo, contadores de bytes entrantes y salientes en los modems de los suscriptores, o en cualquier otro elemento de red como puede ser un

nodo de agregación o un router. Aprovechando el protocolo SNMP se suelen construir sistemas de bases de datos que, mediante un proceso automatizado de extracción, transferencia y carga (ETL) colectan datos por medio de encuestadores (polling) a los elementos de red. Si se pretende almacenar datos históricamente, algunas de las dimensiones como las series de tiempo escalan rápidamente en grandes volúmenes de datos y los motores de bases de datos tradicionalmente monolíticos dejan de ser suficientes. La solución a este problema aparece con las tecnologías distribuidas, como por ejemplo los sistemas de archivos hiper-distribuidos (HDFS) [3], que mediante el algoritmo MapReduce [4] y el paralelismo de hardware, llevan a una nueva concepción en donde las bases de datos y sus motores, de tecnologías distintas, las aplicaciones y un catálogo de interconexiones posibles residen en un mismo clúster. En la figura 1 se muestra esquemáticamente el proceso diseñado y utilizado en este trabajo.

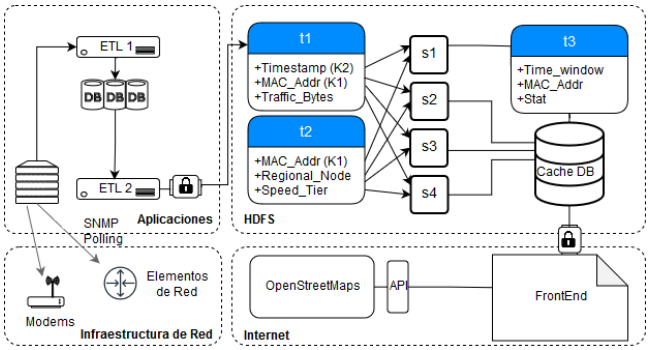

Figura 1. Diagrama esquemático del flujo de datos para una aplicación con datos georreferenciados. Los datos se generan en la Infraestructura de Red y son colectados y administrados mediante procesos ETL que suelen estar segurizados. Utilizando un HDFS se resuelve escala en términos de storage y cómputo mediante otro ETL (t1) e integraciones con otros orígenes (t2). El desarrollo de scripting dentro del cluster (s1, s2, s3, s4) permite la reducción de dimensiones y la creación de una base de datos caché con tablas parametrizables (t3) que son utilizadas por una aplicación web (FrontEnd) integrada mediante APIs a servicios de georreferencia como OpenStreetMaps [5]

II-A. Descripción de los Datos

Con el objetivo de obtener información de la red con distintos niveles de detalle, se integra información proveniente de los sistemas de registración y billing por un lado, y de tráfico en cada modem y elemento de red por otro. El consumo de los usuarios se asocia al tráfico medido en los modems en la red de acceso. Para obtener por ejemplo el consumo mensual, simplemente se suma todo el tráfico correspondiente a un mes de cada modem. Otro indicador útil es el tráfico en la hora pico o consumo promedio en la franja horaria de mayor concurrencia (los domingos de 18 hs a 00 hs [6]). Otros datos que analizamos con el mismo proceso son los estadísticos del tráfico mensual (promedio, mediana, percentiles, máximo, mínimo y desvíos). Por el lado del producto, se analiza cuál es la velocidad media contratada, que consiste en promediar los productos contratados por los clientes en determinado sector.

II-B. Visualización y Navegación

Los datos obtenidos y procesados se muestran en una aplicación web interactiva. Permite conocer el comportamiento de los clientes en distintos niveles de desagregación (total de clientes o por producto, sitio, nodo o zona). Todos los datos corresponden al comportamiento de un cliente promedio. En primer lugar, se muestran diversos mapas con información georreferenciada de los nodos de la red de Fibertel del mes de marzo de 2019 (Figuras 2, 3 y 4). Se puede observar el consumo mensual downstream y upstream, el tráfico promedio mensual downstream y upstream, la cantidad de clientes activos y la velocidad media contratada por los mismos. El nivel de nodo es para el único con el cual contamos con las coordenadas geográficas. Un nivel más agregado es el sitio, y un nivel más desagregado es la zona. Cada variable analizada viene acompañada por su distribución, en la cual se pueden observar los valores más frecuentes, y los rangos con los colores representados en el mapa. Además, al desplazarse sobre el mapa, al pasar por encima de un punto, una etiqueta indica a que nodo corresponde y el valor de la variable, como se puede observar en la Figura 5.

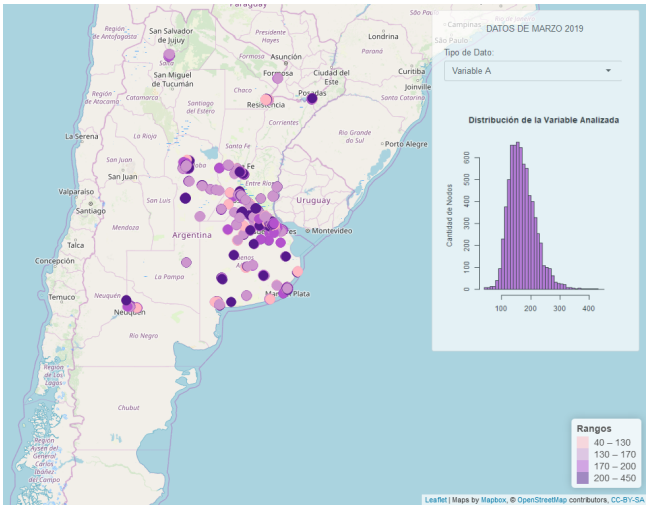

Figura 2. Mapa de la Argentina con información georreferenciada. Cada punto representa a un nodo de la red de Fibertel. Se muestra la distribución de la variable de interés y los rangos de valores con los colores correspondientes.

La aplicación web no solo cuenta con la información del último mes, si no con toda la información histórica disponible, lo que se representa con series temporales e indicadores numéricos, como una forma de visualizar los mismos datos de manera longitudinal, es decir, no a través del espacio sino del tiempo.

Una vez que se cuenta con las series temporales, es de interés poder hacer pronósticos de los valores futuros. A continuación, se analizan diversos métodos predictivos.

II-C. Modelos Predictivos

Las series de consumo y tráfico con las que trabajamos presentan una marcada tendencia creciente y un comportamiento estacional, es decir que los datos experimentan variaciones

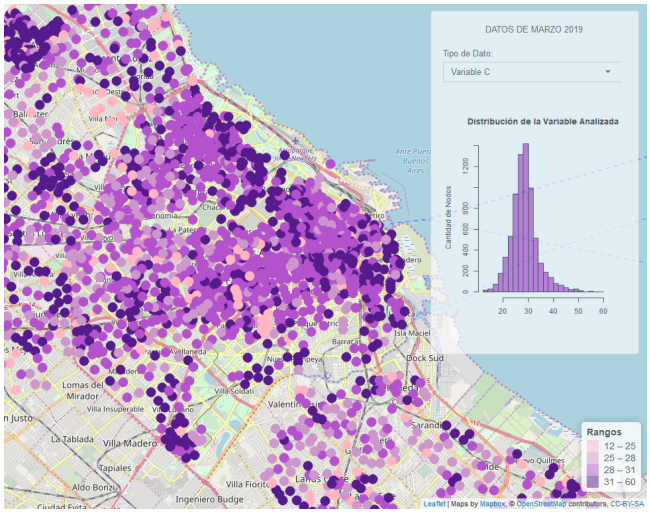

Figura 3. Mapa de la Capital Federal. Se puede apreciar la densidad de nodos en esta zona.

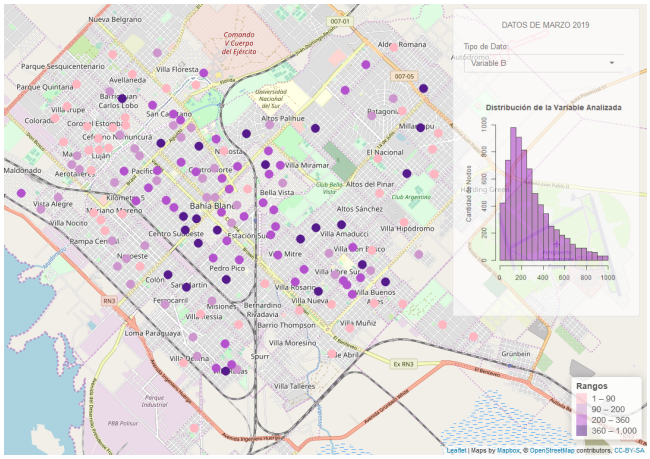

Figura 4. Mapa de Bahía Blanca con sus nodos correspondientes.

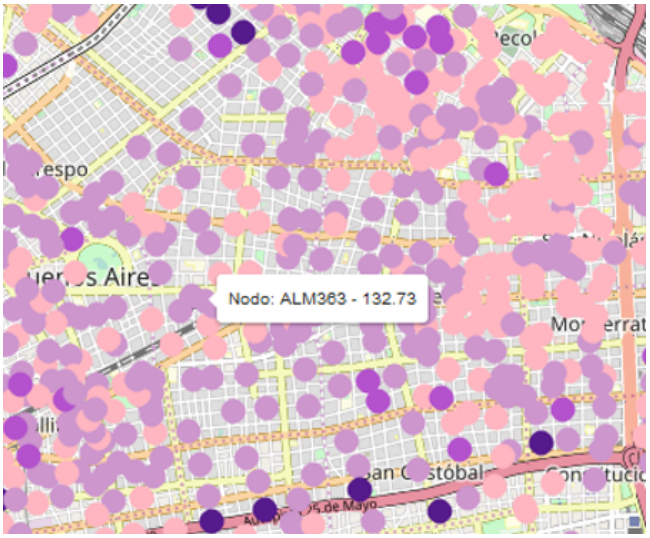

Figura 5. Funcionalidad interactiva: se puede observar a que nodo corresponde cada punto y el valor de la variable que se está analizando.

Figura 6. Gráfico estacional: la serie temporal se gráfica en coordenadas polares para poder apreciar su estacionalidad. Por ejemplo, se ven variaciones en junio y noviembre que se repiten ambos años. Además, la forma en espiral se debe a que la variable es creciente.

regulares y previsibles que se repiten cada año (Figura 6). Es por este motivo que se analizan modelos adecuados para este tipo de series.

Los dos métodos predictivos más utilizados son ARIMA (modelo autorregresivo integrado de media móvil) y Suavizado Exponencial, y presentan acercamientos complementarios al problema [7]. ARIMA se basa en describir las autocorrelaciones en los datos, mientras que el suavizado exponencial se basa en la descripción de la tendencia y la estacionalidad.

En primer lugar, se analizó el modelo ARIMA, para el cual es necesario que la serie sea estacionaria. Esto quiere decir que no debe tener tendencia ni estacionalidad, y la varianza debe ser constante. Para eliminar la tendencia y la estacionalidad, se trabaja con la serie de las diferencias (la componente de integración del modelo) y para lograr varianza constante, se puede trabajar con la variable transformada; lo más común es aplicar logaritmo.

Para eliminar la tendencia se calcula la diferencia de orden 1 de la siguiente manera:

$$y'_t = y_t - y_{t-1}$$

De manera análoga se pueden calcular otros ordenes de diferencias hasta lograr eliminar la tendencia de la serie, en caso de que persista.

Si se sabe que la serie tiene estacionalidad anual, se calcula la diferencia de orden 12 (meses) para eliminarla:

$$y'_t = y_t - y_{t-12}$$

Si quiero diferencias de orden 1 y 12, e y'_t es la serie diferenciada estacionalmente entonces:

$$y''_t = y'_t - y'_{t-1} = y_t - y_{t-1} - y_{t-12} + y_{t-12-1}$$

Una notación útil para trabajar con series diferenciadas es el operador Backshift:

$$By_t = y_{t-1} \ (diferencia\ de\ orden\ 1)$$

$$B^{12}y_t = y_{t-12} \ (diferencia\ de\ orden\ 12)$$

Entonces la serie se escribe:

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1-B)y_t$$

Con esta notación, la serie y''_t mencionada antes queda:

$$y''_t = (1-B)(1-B^{12})y_t$$

Ahora puedo trabajar con la variable obtenida (luego de haber realizado una transformación en caso de que fuera necesario para obtener varianza constante).

El modelo ARIMA está compuesto por un modelo autorregresivo (AR) y un modelo de media móvil (MA) aplicado sobre una serie diferenciada. Los modelos autorregresivos usan una combinación lineal de los valores pasados de la variable. Por lo tanto, un modelo $AR(p)$ se escribe:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

donde ϵ_t es ruido blanco.

En cambio, los modelos de media móvil usan los errores pasados de pronóstico. Un modelo $MA(q)$ se escribe:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

Por lo tanto, un modelo $ARIMA(p,d,q)$ se escribe:

$$y'_t = c + \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} +$$

$$+\theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$$

(con $d = 1$).

Con la notación backshift se ve de esta forma:

$$(1 - \phi_1 B - ... - \phi_p B^p)(1-B)^d =$$

$$= c + (1 + \theta_1 B + ... + \theta_q B^q)\epsilon_t$$

Finalmente, para considerar un ARIMA estacional, se incluyen términos estacionales adicionales $(p,d,q)(P,D,Q)_m$. Un modelo $ARIMA(1,1,1)(1,1,1)_{12}$ sin constante se escribe:

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(1-B)(1-B^{12})y_t =$$

$$= (1+\Theta_1 B)(1-\Theta_1 B^{12})\epsilon_t$$

El método ARIMA estacional requiere de datos de varios años para poder estimar todos los parámetros, y en nuestro caso solo contamos con información desde el 2017. Por lo tanto, recurrimos a otro tipo de modelo que se ajusta bien a series estacionales, el método de suavizado exponencial. Este método tiene la ventaja de que es simple, requiere menos datos para poder predecir y es capaz de adaptarse a los cambios en tendencia y estacionalidad apenas ocurren ya que se enfoca en períodos recientes [8].

Los pronósticos obtenidos mediante este método son promedios ponderados de observaciones pasadas. Los pesos decaen exponencialmente, es por esto que las observaciones más recientes tienen pesos asociados más elevados.

El modelo más sencillo de este tipo es el suavizado exponencial simple, que se escribe:

$$\hat{y}_{t+1|t} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + ...$$

donde α es el parámetro de suavizado $(0 < \alpha < 1)$. Además de α, hay que elegir un valor inicial para $t = 1$; lo denotamos n_0.

Entonces el pronóstico a tiempo $t+1$ es un promedio ponderado entre la observación más reciente y el pronóstico previo:

$$\hat{y}_{t+1|t} = \alpha y_t + \alpha(1-\alpha)\hat{y}_{t|t-1}$$

$$\hat{y}_{t+1|t} = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j y_{t-j} + (1-\alpha)^t n_0$$

Otra notación útil es la forma de componentes. Para el caso simple, el nivel es la única componente; entonces:

$$Pronóstico \quad \hat{y}_{t+h|t} = n_t$$

$$Nivel \quad n_t = \alpha y_t + (1-\alpha)n_{t-1}$$

donde n_t es el nivel estimado a tiempo t, y α es el parámetro de suavizado para el nivel $(0 < \alpha < 1)$.

A medida que avanzamos, los métodos se van complejizando. El siguiente es una extensión del modelo simple que permite trabajar con datos con tendencia lineal (suavizado exponencial lineal de Holt); incluye una ecuación de pronóstico y dos ecuaciones de suavizado (nivel y tendencia):

$$Pronóstico \quad \hat{y}_{t+h|t} = n_t + hb_t$$

$$Nivel \quad n_t = \alpha y_t + (1-\alpha)(n_{t-1} + b_{t-1})$$

$$Tendencia \quad b_t = \beta(n_t - n_{t-1}) + (1-\beta)b_{t-1}$$

donde b_t es la pendiente (tendencia) estimada a tiempo t, y β es el parámetro de suavizado para la tendencia $(0 < \beta < 1)$.

Hasta acá, podemos modelar series que mantienen la tendencia constante indefinidamente. Esto no suele ocurrir, sobre todo para horizontes muy extensos. Por lo tanto, se introduce un parámetro que amortigua la tendencia (suavizado exponencial de tendencia amortiguada):

$$Pronóstico \quad \hat{y}_{t+h|t} = n_t + (\delta + \delta^2 + ... + \delta^h)b_t$$

$$Nivel \quad n_t = \alpha y_t + (1-\alpha)(n_{t-1} + \delta b_{t-1})$$

$$Tendencia \quad b_t = \beta(n_t - n_{t-1}) + (1-\beta)\delta b_{t-1}$$

donde δ es el parámetro de amortiguación $(0 < \delta < 1)$.

Por último, el método de Holt – Winters extiende lo visto hasta el momento para poder trabajar con series estacionales. Se agrega una tercera ecuación de suavizado, la responsable del componente estacional (e_t). Tiene dos variantes: el método aditivo y el multiplicativo. El primero se utiliza cuando las variaciones estacionales son aproximadamente constantes a lo largo de la serie, mientras que el segundo se utiliza cuando

dichas variaciones cambian proporcionalmente al nivel de la serie.

Método aditivo:

$$Pronóstico \quad \hat{y}_{t+h|t} = n_t + hb_t + e_{t+h-m(k+1)}$$

$$Nivel \quad n_t = \alpha(y_t - e_{t-m} + (1-\alpha)(n_{t-1} + b_{t-1})$$

$$Tendencia \quad b_t = \beta(n_t - n_{t-1}) + (1-\beta)b_{t-1}$$

$$Estacionalidad \quad e_t = \gamma(y_t - n_{t-1} - b_{t-1}) + (1-\gamma)e_{t-m}$$

donde γ es el parámetro de estacionalidad $(0 < \gamma < 1)$, $m = 12$ para datos mensuales, y $k = \lfloor (h-1)/k \rfloor$ para asegurar que las estimaciones de los índices estacionales sean del último año de la muestra.

Método multiplicativo:

$$Pronóstico \quad \hat{y}_{t+h|t} = (n_t + hb_t)e_{t+h-m(k+1)}$$

$$Nivel \quad n_t = \alpha\frac{y_t}{e_{t-m}} + (1-\alpha)(n_{t-1} + b_{t-1})$$

$$Tendencia \quad b_t = \beta(n_t - n_{t-1}) + (1-\beta)b_{t-1}$$

$$Estacionalidad \quad e_t = \gamma\frac{y_t}{n_{t-1} + b_{t-1}} + (1-\gamma)e_{t-m}$$

A esos dos métodos se les puede agregar amortiguación como vimos anteriormente. Por ejemplo, el método multiplicativo con amortiguación se escribe:

$$Pronóstico \quad \hat{y}_{t+h|t} = [n_t + (\delta + \delta^2 + ... + \delta^h)b_t]e_{t+h-m(k+1)}$$

$$Nivel \quad n_t = \alpha\frac{y_t}{e_{t-m}} + (1-\alpha)(n_{t-1} + \delta b_{t-1})$$

$$Tendencia \quad b_t = \beta(n_t - n_{t-1}) + (1-\beta)\delta b_{t-1}$$

$$Estacionalidad \quad e_t = \gamma\frac{y_t}{n_{t-1} + \delta b_{t-1}} + (1-\gamma)e_{t-m}$$

III. Resultados

Se compararon todos los modelos de suavizado exponencial y, además, se aplicaron distintas transformaciones a los datos de entrada:

- (sin transformar las variables)
- Logaritmo
- Raíz cuadrada
- Función logística

Para elegir el mejor modelo en cada caso, se busca aquel que cumpla los supuestos de residuos independientes y con distribución normal, que optimice ciertas métricas de ajuste y que tenga menor intervalo de confianza. En todos los casos, el tipo de modelo que siempre ajusta mejor es el Holt - Winters aditivo, con distintas transformaciones de la variable. Como ejemplo, en la Figura 7 se muestra la predicción de una de nuestras variables de análisis.

Lo hecho hasta el momento permitiría en un futuro automatizar la fase de predicción y poder predecir las diferentes variables de análisis en todos los niveles (sitio, nodo, zona). Para esto, es necesario cuantificar todos los criterios de selección para que puedan ser comparados de manera automática.

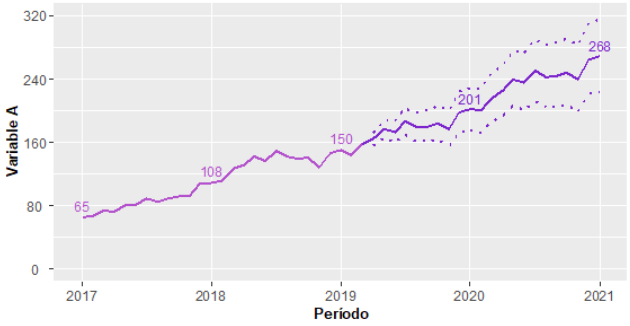

Figura 7. Serie temporal de una variable de interés. Se muestran los valores históricos más los valores pronosticados con el correspondiente intervalo de confianza.

Aplicación a la Industria

Conocer en detalle los datos de la red es sumamente importante porque nos permitiría destinar de manera eficiente los recursos de los cuales dispone la compañía. Por ejemplo, enfocar las obras de dimensionamiento más urgentes a zonas específicas de la red, o realizar campañas dirigidas a ciertas zonas geográficos o grupos de clientes.

Existen otras industrias bien distintas a la de servicios digitales que también pueden hacer uso de las tecnologías y métodos presentados en este trabajo. Por ejemplo, las redes de suministro eléctrico tienen variables similares: los tipos de tarifas, el consumo, la estacionalidad. Las redes de distribución de gas, las de aguas sanitarias y las de autopistas viales también tienen la misma suerte de datos.

Por un lado, los métodos de pronóstico son útiles para hacer planificación de inversiones de mediano y largo plazo con mucho nivel de detalle, y por otro, los métodos de visualización y navegación de datos son útiles para el mantenimiento proactivo y preventivo de fallas.

IV. Conclusiones

Este trabajo tiene como novedad la presentación de datos georreferenciados de uso de internet en Argentina para el período 2017-2019. La visualización sobre mapas es muy útil para condensar mucha información en una sola imagen. De hecho, la navegación de distintas variables se vuelve muy práctica.

Estudiar las series de tiempo y su naturaleza permite aplicar diversas metodologías para hacer pronósticos. Hemos visto que aparecen relaciones de compromiso entre cantidad de datos disponibles y resultados. En particular, el método ARIMA estacional no solo requiere muchos datos [9], si no también mucha historia. Por eso se elige un método más simple y práctico, como el suavizado exponencial.

Por último, la estructura de datos definida en este trabajo, utilizando arquitecturas de código abierto, hace a la aplicación web exportable a otras industrias.

REFERENCIAS

[1] Bonfim, Michel S. and Dias, Kelvin L. and Fernandes, Stenio F. L., Integrated NFV/SDN Architectures: A Systematic Literature Review, ACM Comput. Surv., New York, NY, USA, February 2019.

[2] Weldon, Marcus K., The Future X Network: A Bell Labs Perspective, CRC Press, Inc., Boca Raton, FL, USA, 2015.

[3] , Shvachko, Konstantin and Kuang, Hairong and Radia, Sanjay and Chansler, Robert, The Hadoop Distributed File System, Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), IEEE Computer Society, Washington, DC, USA, 2010.

[4] Dean, Jeffrey and Ghemawat, Sanjay, MapReduce: Simplified Data Processing on Large Clusters, Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, USENIX Association, Berkeley, CA, USA, 2004.

[5] OpenStreetMap contributors, Planet dump retrieved from https://planet.osm.org, https://www.openstreetmap.org, 2017.

[6] Righetti, Claudio and Gibellini, Emilia and De Arca, Florencia and Carreño Romano, Carlos Germán and Fiorenzo, Mariela and Ochoa, Fernando Rodrigo and Carro, Gabriel, Network Capacity and Machine Learning, A Technical Paper prepared for SCTE/ISBE, Cable-Tec Expo 2017, Denver, October 2017.

[7] Rob J. Hyndman and George Athanasopoulos, Forecasting: Principles and Practice, online book, 5/9/2018.

[8] Paul Goodwin, The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong, ResearchGate, 3/6/2014.

[9] Carreno Romano, Carlos Germán and Clivio, Natalia and Righetti, Claudio, Sizing Techniques applied to Network Capacity Planning, ARGENCON, 2018.