INAP Campus Virtual

Big Data: Haciendo hablar los datos

INAP Instituto Nacional de la Administración Pública

Secretaría de Gestión y Empleo Público

Jefatura de Gabinete de Ministros
Argentina

Índice

Introducción

Hoy estamos atravesados por los datos y continuamente los estamos generando. La cantidad de datos que hoy generamos es enorme: no solo vienen de cuentas de correo electrónico, WhatsApp, Facebook, Twitter, fotos digitales, GPS y videos, sino también de sensores de climatológicos, datos socioeconómicos, fotos satelitales, etc. Para tener una somera idea, por minuto se suben a YouTube más de 400 horas de videos.

Hoy se toman y analizan datos de deportes, salud, casas inteligentes, venta de supermercados y el posicionamiento en góndola, consumo de productos y servicios, publicidad efectiva, etc. Por ejemplo, en el caso de los deportes, hay equipos de fútbol que hacen seguimiento de la pelota y de los 22 jugadores para mejorar el rendimiento. También, en el caso del automovilismo de alta competencia, los autos tienen sensores que reportan en tiempo real el desempeño del auto y del piloto. Aun en el ciclismo se usan sensores que describen características fisiológicas de los ciclistas. En lo relacionado con el consumo de productos y servicios, se analizan los gustos para presentarle al cliente nuevos productos que puedan ser de su interés. En el campo de la salud, hoy existen dispositivos que miden el ritmo cardíaco, el movimiento de los ojos y la actividad cerebral para poder corregir disfunciones, como también para diagnosticar enfermedades.

Pero ¿de qué hablamos cuando hablamos de big data?

Big data (‘macrodatos’ o ‘inteligencia de datos’ en español) es un concepto que se ha puesto de moda y tiene que ver con poder analizar de una manera útil los datos disponibles; con la habilidad para modelar y analizar datos, independientemente de cuán grandes sean.

Cuando uno tiene grandes volúmenes de datos, existen desafíos teóricos y computacionales. Por ejemplo, ante los datos que produce el acelerador de partículas de la Organización Europea para la Investigación Nuclear (conocida por la sigla en francés CERN), uno está interesado en analizarlos en tiempo real. Sin embargo, en general, en los casos de big data con que nos topamos, el tamaño de los datos no es monstruoso para las capacidades computacionales con las que solemos contar. El desafío, por lo tanto, pasa por ser capaces de «hacer hablar a los datos» y, a partir de ahí, por proponer nuevos caminos y cambios estratégicos.

¿Cómo nos afectan los macrodatos en las políticas públicas?

Hoy los países tienden a tomar todas sus medidas basadas en datos. Es decir, se busca que la toma de decisiones sea transparente basada en la información disponible. Y el desafío es analizar correctamente los datos para poder proponer nuevas políticas que mejoren y simplifiquen la vida de los ciudadanos.

Ejemplos:

- Ciudades inteligentes: crecimiento del parque automotor, transporte, calidad del aire, distribución geográfica de escuelas, hospitales, servicios públicos de atención, etc.
- Tarjeta sube: hoy se cuenta con la información de en qué momento un pasajero subió a determinado medio de transporte público y el posicionamiento del mismo. ¿Es posible diseñar mejoras en los medios de transporte a partir de esta información?

- Fotografías satelitales para dar mayor previsibilidad a ciertos sectores de la economía
- Enfermedades: control de brote de epidemias.
- Mapa del delito y políticas de seguridad.
- Nuevos mecanismos de otorgamiento de préstamos sociales desde el sector público.
- Modelar el impacto de cierta política impositiva (antes de ser aplicada).
- Nuevas maneras de tipificar/clasificar a las empresas.

DATOS Y DESAFÍOS

Se puede considerar que hay cinco aspectos importantes sobre los datos:

- **Generación y captura:** ¿qué queremos medir?, ¿cuál es la tecnología adecuada para medirlo?
- **Protección de datos:** Los datos tienen que estar protegidos contra pérdidas y amenazas de corrupción.
- **Apertura de datos (open data):** está claro que, si se logra que los datos estén disponibles para cualquier persona, esto es extremadamente favorable. Por un lado, muchísima más gente puede pensar en nuevos modelos y formas de analizar estos datos llegando antes a mejores soluciones, disminuyendo así el error de un posible mal análisis realizado por unas pocas personas. Y, por otro lado, otro de los grandes beneficios es que permite mostrar transparencia en la toma de decisiones: todas las decisiones están debidamente justificadas. Además permite medir públicamente indicadores de buena gestión a partir del cumplimiento o metas basada en datos reales.
- **Limpieza de los datos:** muchas veces los datos crudos (raw data) cuentan con datos espurios, datos faltantes o simplemente no tenemos una nomenclatura en común en la información obtenida. Por eso, muchas veces se debe realizar un arduo trabajo para llevarlos a un formato adecuado para poder luego analizarlos.

- **Análisis de los datos:** ¿cómo hacer que los datos hablen? Esto se puede considerar un arte, pero, como veremos, existen técnicas analíticas que nos ayudan a avanzar sobre esta pregunta.

ANÁLISIS DE DATOS

A continuación, presentaremos las metodologías de exploración y análisis de datos basándonos en ejemplos concretos de la administración pública:

- **Salud pública:** distribución geográfica de hospitales/centro de atención primaria, tiempos de espera.

- **Electricidad:** patrones de consumo; políticas de consumo que incentiven a consumir menos en horarios de máxima demanda (nuevas tarifas).

Salud pública

En salud pública, contamos con mucha información que continuamente es anali-zada, pero hay mucho por hacer. Entre los desafíos que se plantean analizando datos de pacientes se encuentran: avanzar hacia una medicina personalizada, mejorar los métodos diagnósticos tempranos, y mejorar la capacidad de respues-ta del sistema de salud pública frente a los diferentes cambios en la sociedad.

La epidemiología es la disciplina que se ocupa del control y el seguimiento de los factores relacionados con la salud y las enfermedades existentes en nuestra población. Son muchas las aristas que hay que analizar al momento de determi-nar si las políticas en término de salud pública son apropiadas. El Ministerio de Salud y Desarrollo Social de la Nación no solo cuenta con información de defun-ciones, nacimientos y casuística de enfermedades por provincia, sino también con información que nos permite entender qué centros de salud necesitan una mayor asistencia económica del Estado.

Tipos de datos, diagrama de barra e histograma (distribuciones)

En esta etapa, prestaremos especial atención al tipo de dato que queremos analizar.

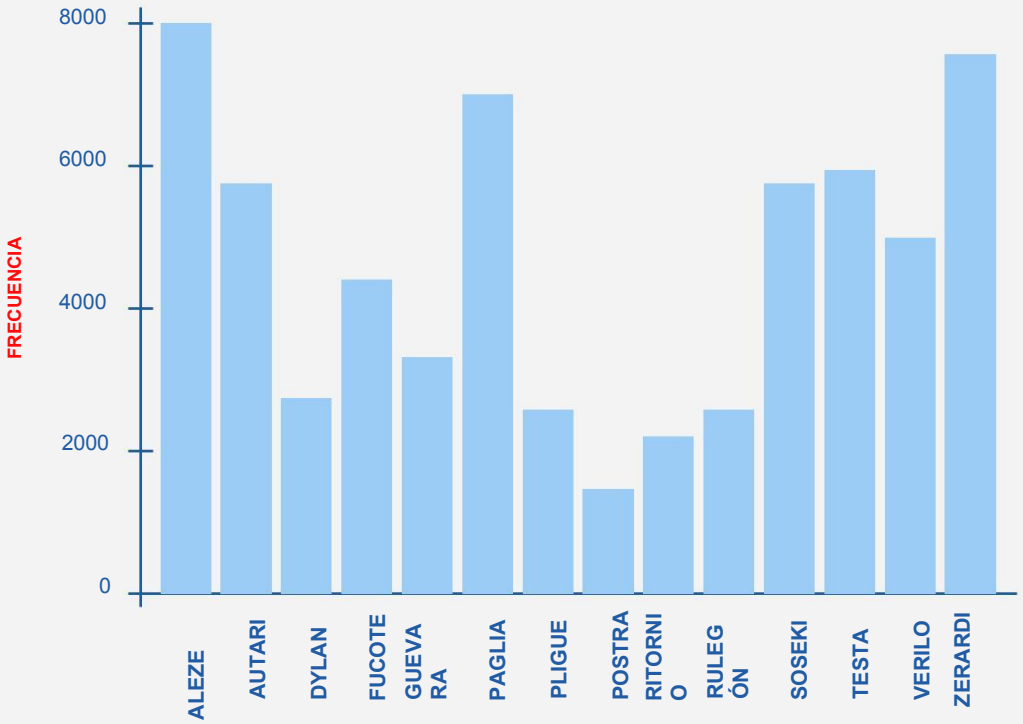
La Argentina cuenta con hospitales generales de agudos, hospitales generales de niños, hospitales especializados, centros de salud, centros médicos barriales, unidades de pronta atención.

Supongamos que tenemos la siguiente información de cada paciente atendido en el país:

1- Nombre
2- Domicilio
3- Centro de atención
4- Entró por guardia
5- día
6- hora de llegada
7- hora en que fue atendido .

Centro de atención	Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora de ser atendido
H. Fucote	Juan Pérez	Arenales 843	07/09/18	Médica	07:32	08:04
H. Fucote	Romina Paz	Malabia 1820	07/09/18	Emergencia	09:35	09:46
H. Autari	Pedro Ast	Zapiola 2232	07/09/18	Emergencia	07:32	08:34
H. Autari	graciela Fort	Castillo 156	07/09/18	Médica	09:45	10:26
H. Aleze	Lía Sope	Bogotá 1566	07/09/18	Emergencia	05:56	06:02
H. Aleze	Carlos Seguí	Caracas 578	07/09/18	Emergencia	10:35	11:12

Estudiamos la variable Centro de atención, ¿cómo representar esta información gráficamente?



¿Cómo resumir la información en una tabla?

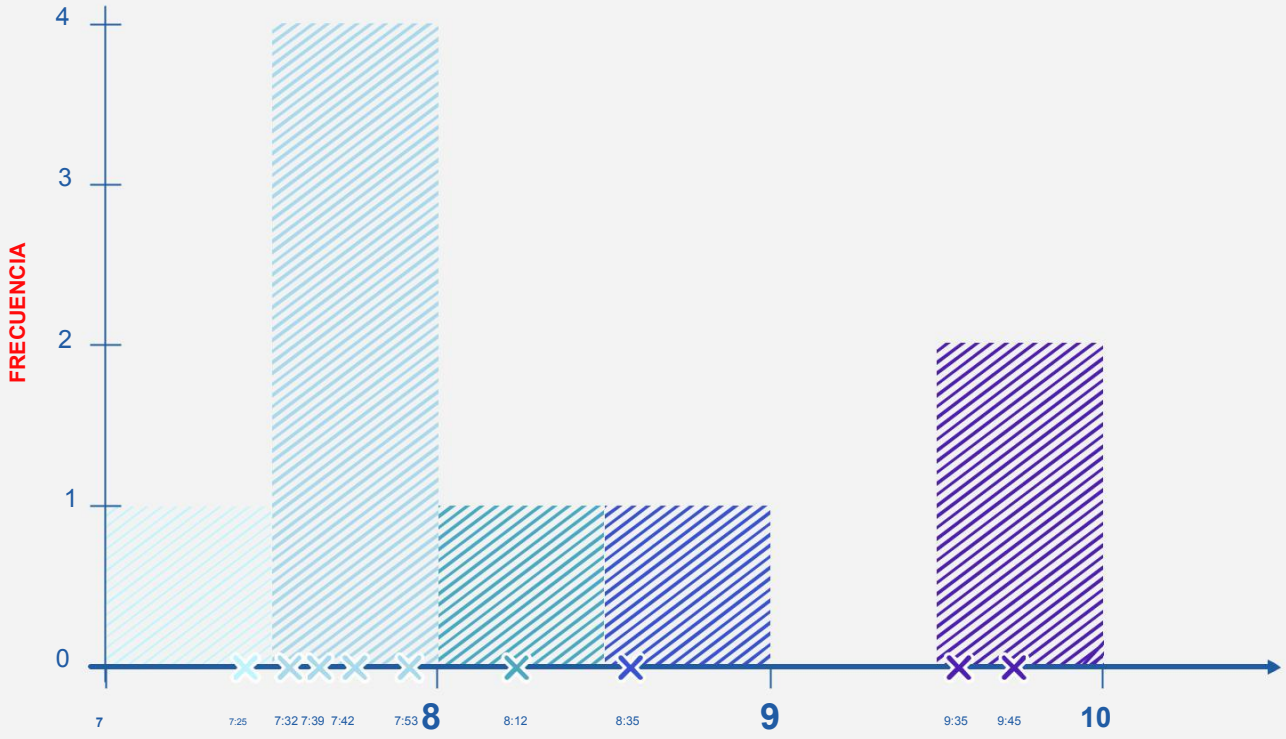
Hospital	Frecuencia
Aleze	8000
Autari	5600
AutGuevaraari	3200
Dylan	2500
Fucote	4300
Paglia	6500
Pligue	1200
Postra	2400
Postra	1900
Rulegón	2400
Soseki	5200
Testa	5300
Verilo	4300
Zerardi	7230

La variable Centro de Atención es una variable categórica. Cada uno de los valores que toma esta variable es una categoría (hospitales generales de agudos, hospitales generales de niños, hospitales especializados, centros de salud, centros médicos barriales, unidades de pronta atención). Lo mismo le sucede a la variable Guardia, que puede tomar solamente las categorías Sí o No. Si queremos representar gráficamente a este tipo de variables, hacemos un diagrama de barra o de torta. Y, si queremos resumir la información en una tabla, hacemos lo de arriba.

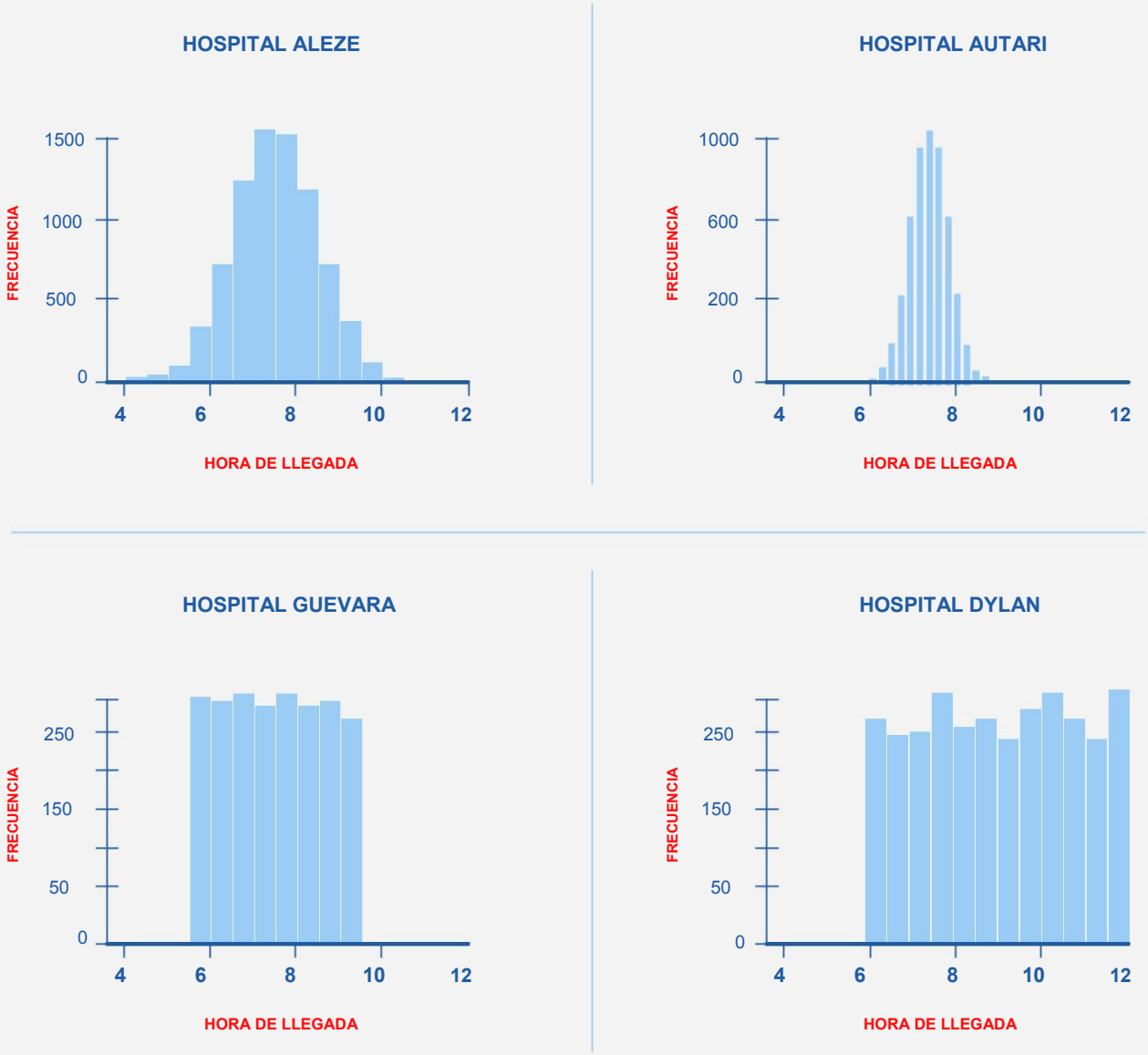
En cambio, la hora en que llegó el paciente y la hora en que fue atendido son dos variables numéricas (no son categorías). Podemos calcular a qué hora llegan en promedio los pacientes y a qué hora son atendidos (en promedio, llegan a las 9:45 y son atendidos a las 10:32), pero como veremos estamos resumiendo mucho la información y probablemente nos estemos perdiendo información relevante.

¿Cómo representar gráficamente la información de la hora en que llegan los pacientes? ¿Qué pasa si convertimos nuestra variable numérica en una categórica?

Si un paciente llegó a las 8:32, decimos que llegó entre las 8:30 y las 8:35 (categoría 8:30-8:35); si llegó a las 7:24, decimos que llegó entre las 7:20 y las 7:25 (categoría 7:20-7:25), etc. Y ahora, al igual que antes, tenemos la frecuencia (cuántos datos hay) en cada una de las categorías. Por lo tanto, podemos hacer un diagrama de barras. En este caso, el gráfico se llama histograma. Su forma varía con el ancho de las clases o categorías.


FRECUENCIA
4
3
2
1
0
7
7:25
7:32
7:39
7:42
7:53
8
8:12
8:35
9
9:35
9:45
10

El histograma es la mejor técnica gráfica para entender cómo están distribuidos (forma del histograma) los datos. En este gráfico, perdemos muy poca información que, en general, no es muy importante: el nombre del paciente y la hora exacta en la que llegó (para 8:32, decimos entre 8:30 y 8:35). La forma es muy relevante y nos ayuda a entender los fenómenos subyacentes.

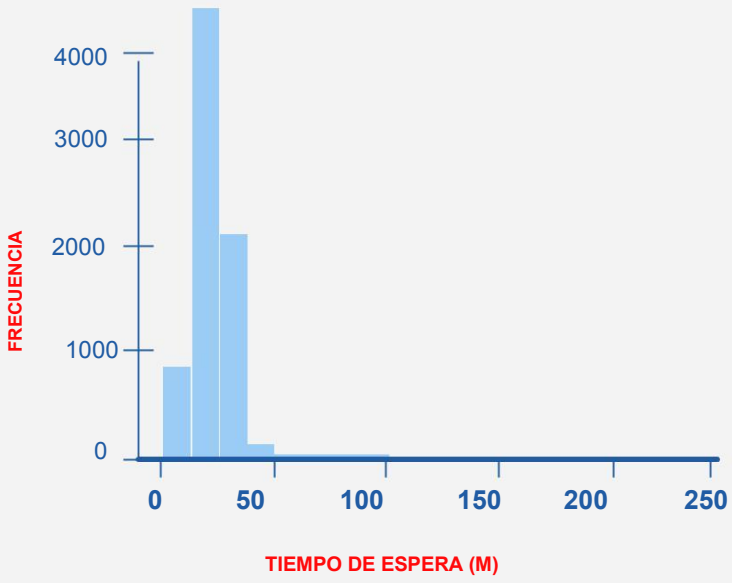


Pero más interesante es pararse en los zapatos del paciente y estudiar cuánto tardó en ser atendido. ¿Cómo es la distribución de tiempos de espera a ser atendido?

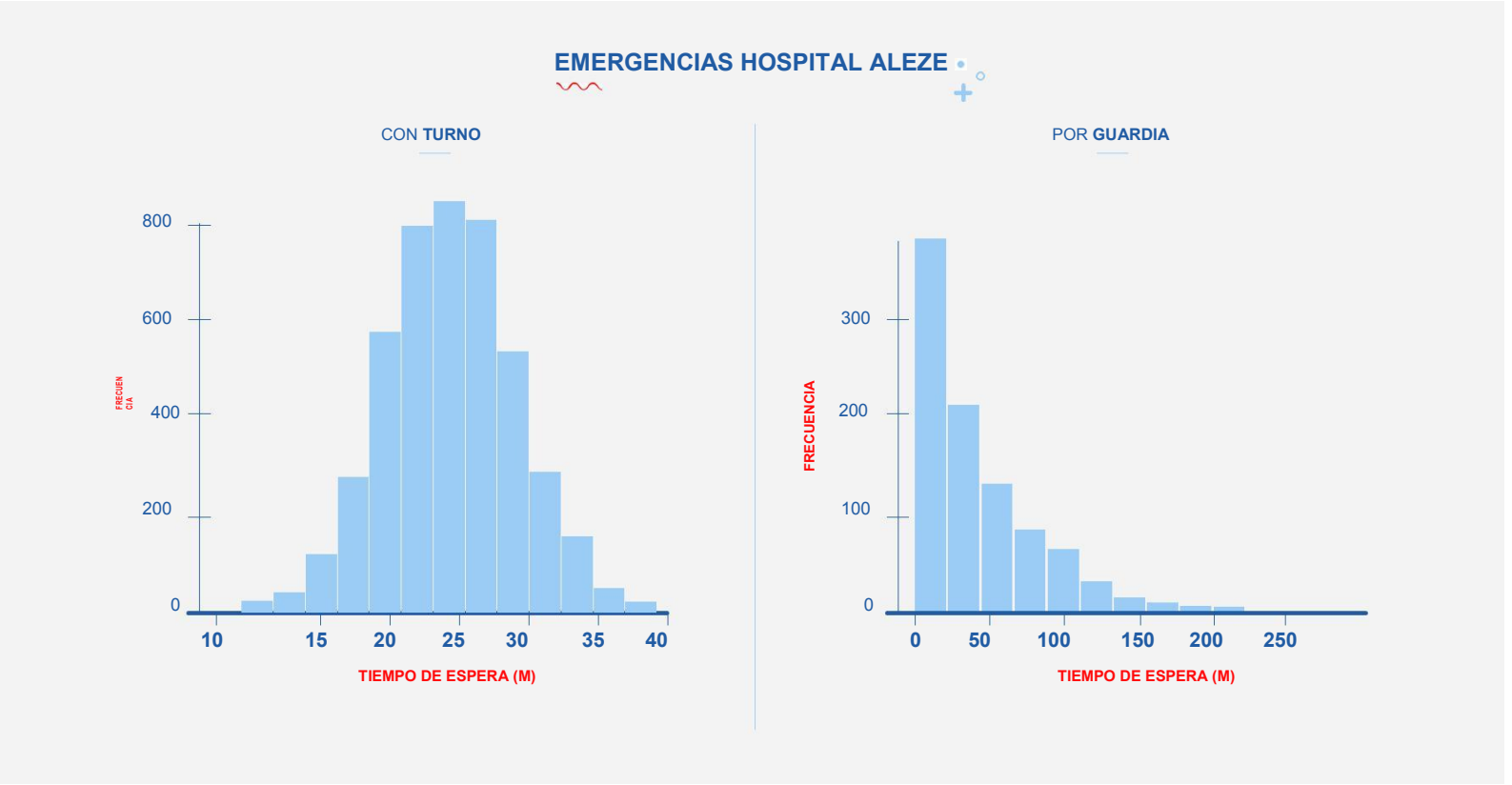
Definimos una variable T.

T = hora atendido − hora de llegada

Supongamos que un paciente llega a las 9:05 h y lo atienden a las 9:55 h. En ese caso, nuestra variable T sería 9:50 h menos 9:05 h, que da un total de 50 minutos de espera.



Separemos los que entran por guardia y los que van con turno:



En la guardia, hay gente a la que atienden muy rápido y otra que espera mucho. Esto se puede deber a que llegó un paciente muy grave y el resto tiene que esperar, o a que se llenó de gente la guardia y entonces los últimos tienen que esperar mucho. En cambio, cuando los pacientes llegan con turno más o menos se respeta el horario.

Supongamos que además contamos con la siguiente información de cada centro de asistencia médica:

- Centro de atención
- Infraestructura [m2 construidos]
- Número de médicos en guardia

Hospital	Infraestructura guardia (m2)	Números de médicos en guardia
Aleze	350	9
Autari	550	15
Guevara	460	8

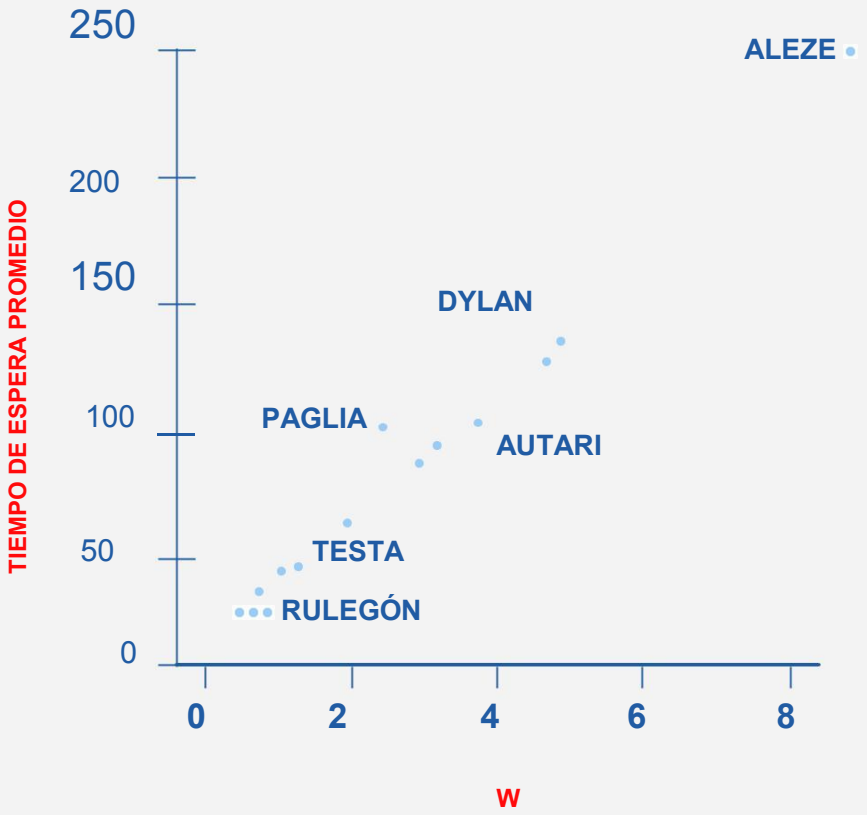
Ahora sí estudiaremos los promedios para comparar hospitales. En particular, queremos entender cuáles son los hospitales que tienen un gran tiempo de espera.

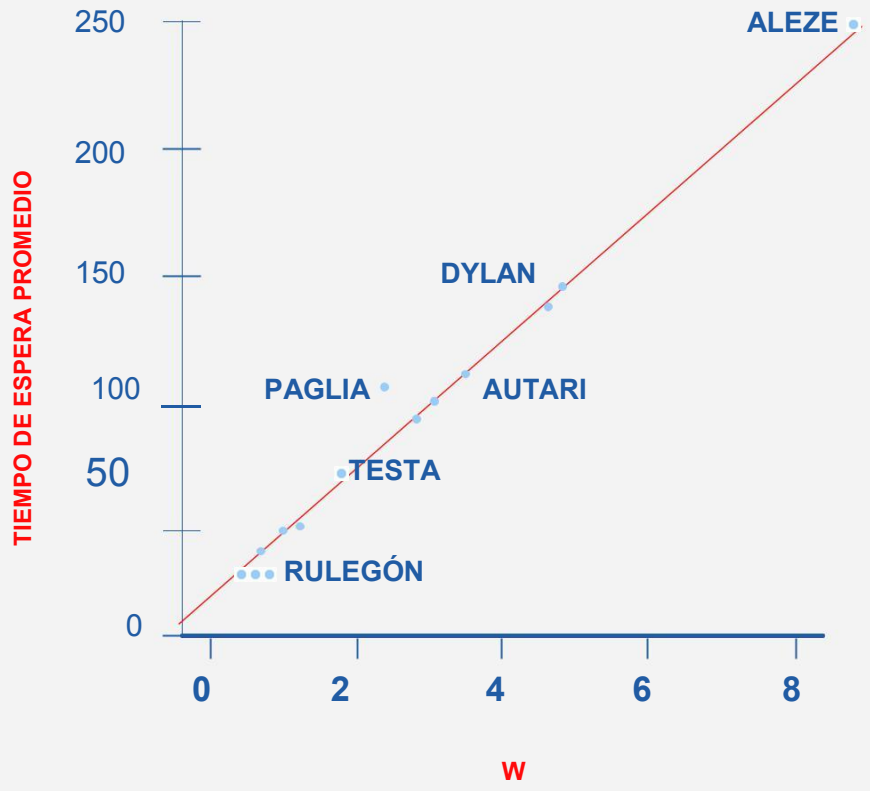
Definimos la variable W = número promedio de pacientes que se presentan por día/número de doctores.

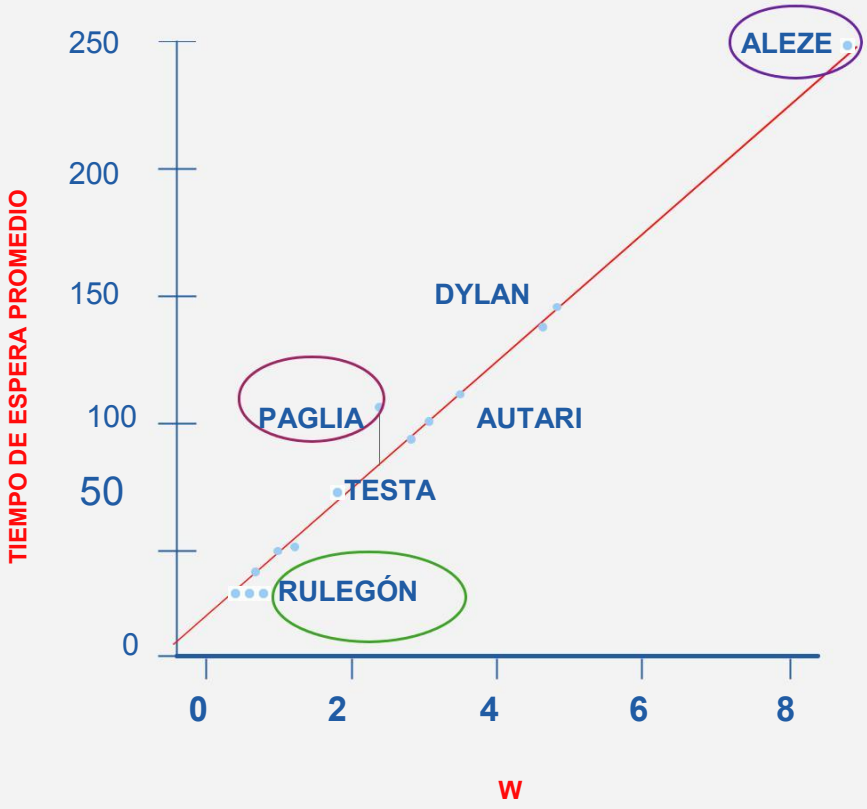
Construimos la base:

Hospital	W	Tiempo de espera promedio (min)
Aleze	8,1	253
Autari	3,6	92
Guevara	1,4	39

Y estudiamos la relación entre W y el tiempo promedio de espera definido como T.


250
200
150
100
50
0
TIEMPO DE ESPERA PROMEDIO
ALEZE
DYLAN
PAGLIA
AUTARI
TESTA
RULEGÓN
0
2
4
6
8
W


250
200
150
100
50
0
TIEMPO DE ESPERA PROMEDIO
ALEZE
DYLAN
PAGLIA
AUTARI
TESTA
RULEGÓN
0
2
4
6
8
W



Dos cosas importantes se pueden observar en el gráfico:

- Por un lado, a mayor W mayor es T, lo cual era de esperar. Es decir, si vienen muchos pacientes para la cantidad de médicos que hay, entonces los pacientes tendrán que esperar un largo rato en ser atendidos.
- Por otro lado, se observan dos hospitales atípicos uno que parecería ser muy eficiente y otro muy poco eficiente.

Hay que tener cuidado con la interpretación de estos dos hospitales atípicos porque podrían existir otras variables que no estamos teniendo en cuenta que expliquen esta diferencia. Por ejemplo, el hospital puede contar con menos o más camillas destinadas a la atención de guardia... Lo que es seguro es que vale la pena indagar cuál es el motivo que explica este alejamiento del patrón de comportamiento.

En el hospital Aleze, los pacientes tienen que esperar muchísimo para ser atendidos: en promedio 250 minutos. Supongamos que queremos mejorar esta situación.

Opciones:

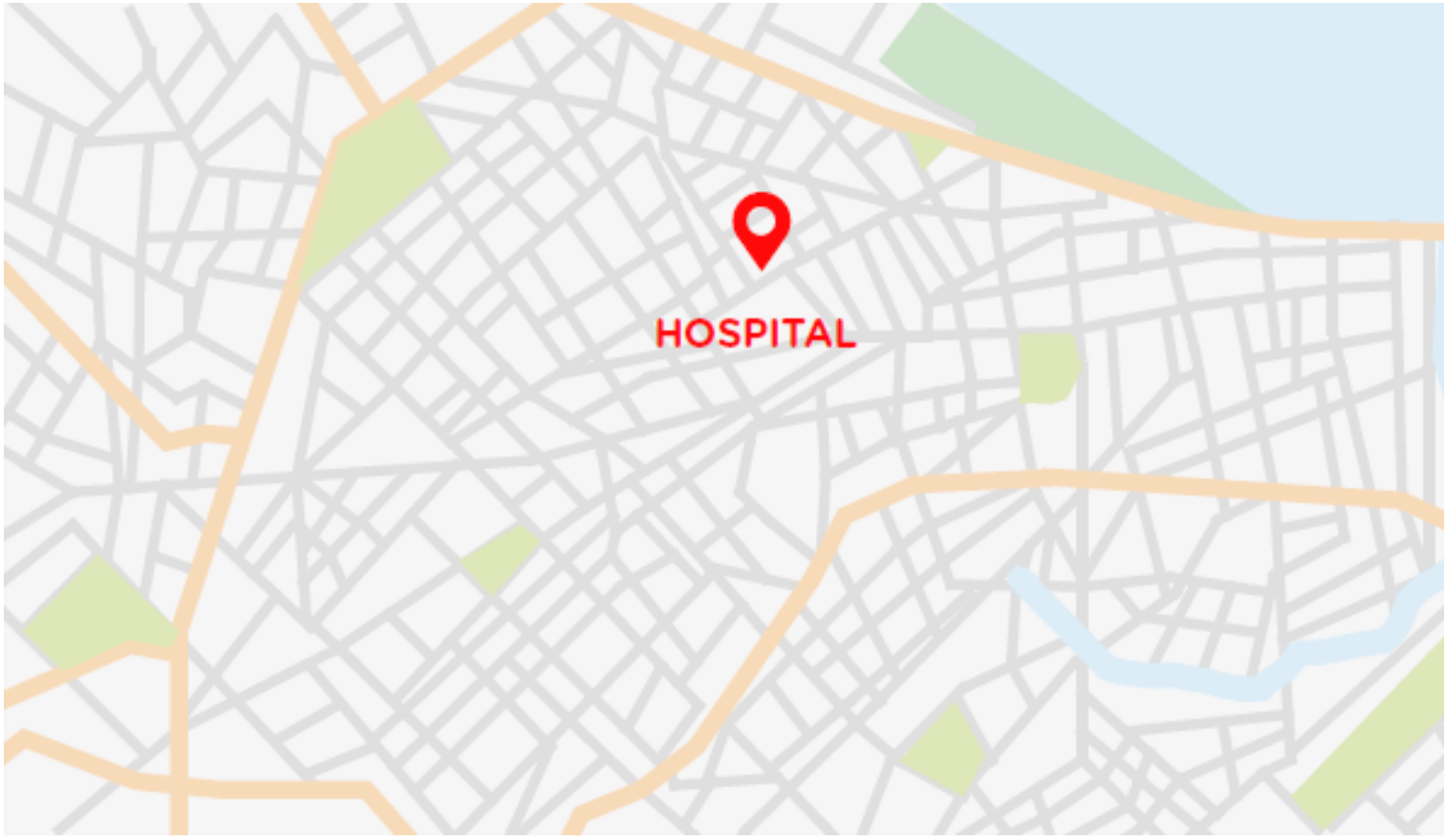
- **Aumentamos el denominador de W:** aumentamos el número de doctores y probablemente la infraestructura (más salas de guardia).

- **Disminuimos el numerador de W:** agregamos una centro de emergencias en una zona aledaña.

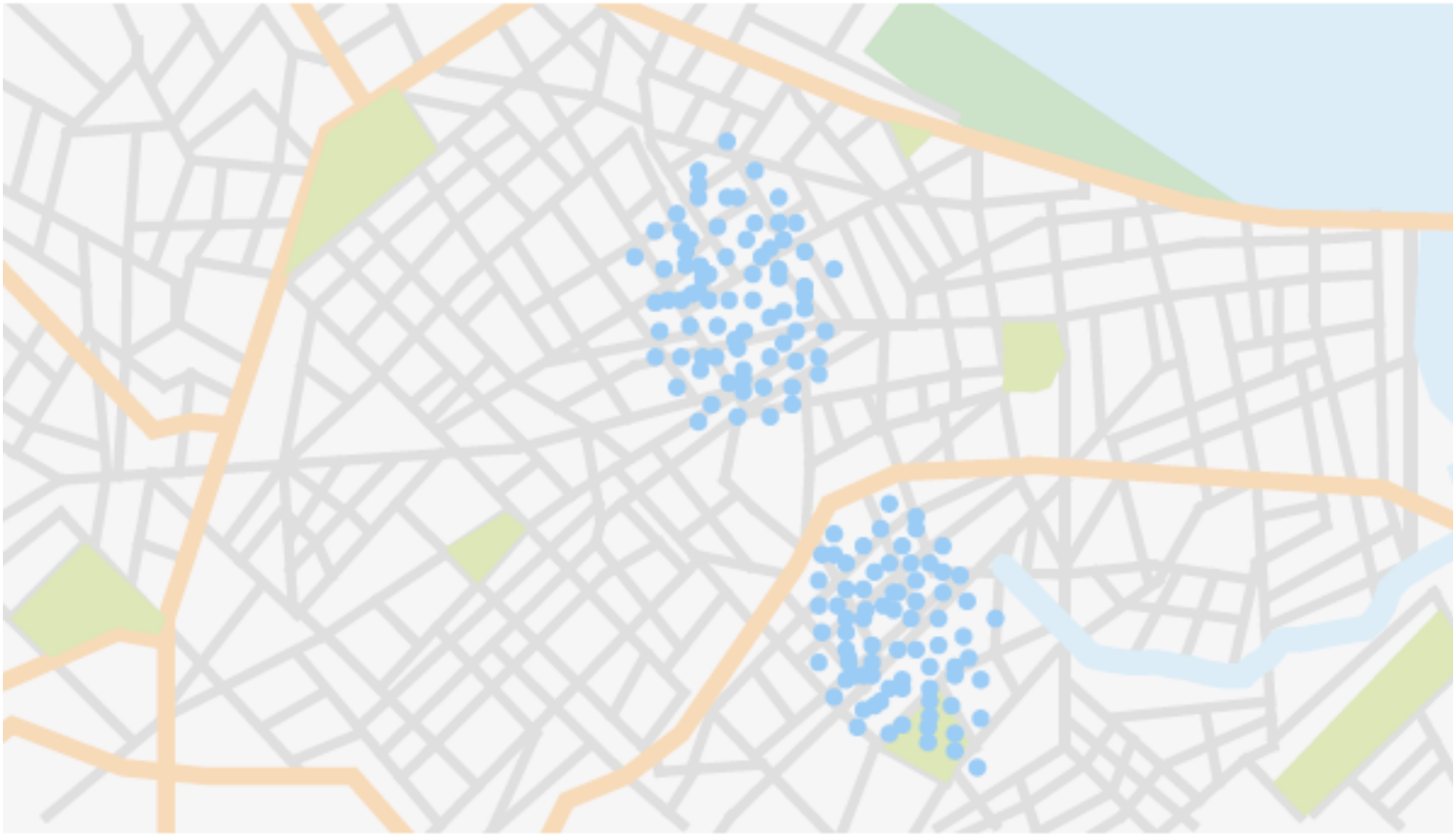
Supongamos que decidimos ir por la opción 2. La pregunta que naturalmente surge es: ¿dónde es recomendable construir el nuevo centro de emergencias?

Si revisamos la información disponible, nos damos cuenta de que tenemos el domicilio de todas las personas atendidas en el hospital Aleze. Por lo tanto, podemos representar cada uno de estos domicilios en un mapa mediante puntos.

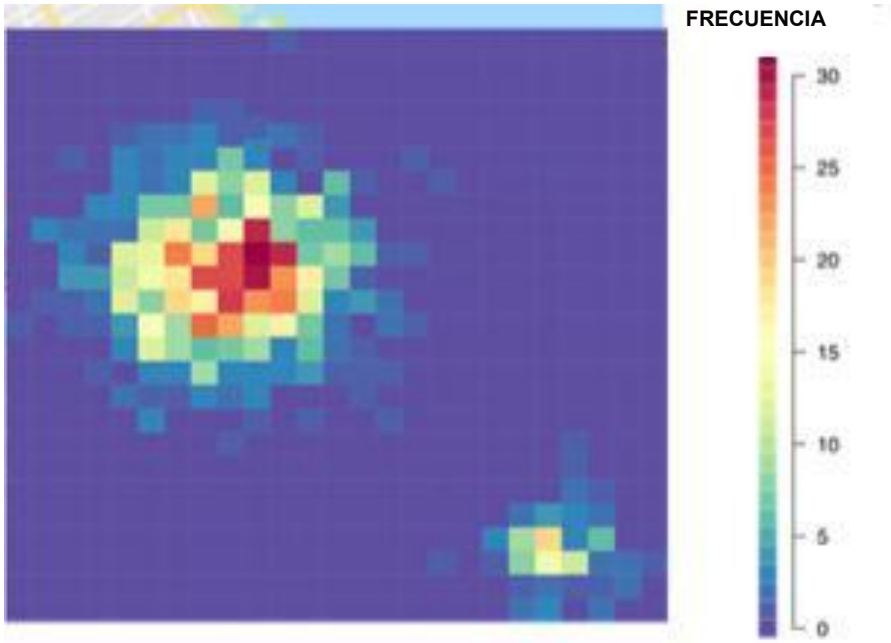
HOSPITAL ALEZE					
Nombre	Domicilio	Día	Consulta	Hora de llegada	Hora de ser Atendido
Lara Menzi	Larralde 382	07/09/2018	Emergencia	05:56	06:02
Mario Parra	Cabildo 54	07/09/2018	Emergencia	06:46	07:24



A partir de este mapa, podemos fijar el centroide del grupo de domicilios lejanos, e intentar construir el nuevo centro médico cerca de este lugar.



Los domicilios también nos pueden servir para mostrar el radio real de cobertura geográfica real de cada centro de salud. Podemos hacer lo mismo que hicimos en el histograma, pero ahora en una grilla de un mapa. Por ejemplo, definimos regiones de 10 manzanas en el mapa y nos fijamos la frecuencia de cada región (número de pacientes que fueron al hospital A en el año cuyo domicilio pertenece a cada uno de las regiones de 10 manzanas).

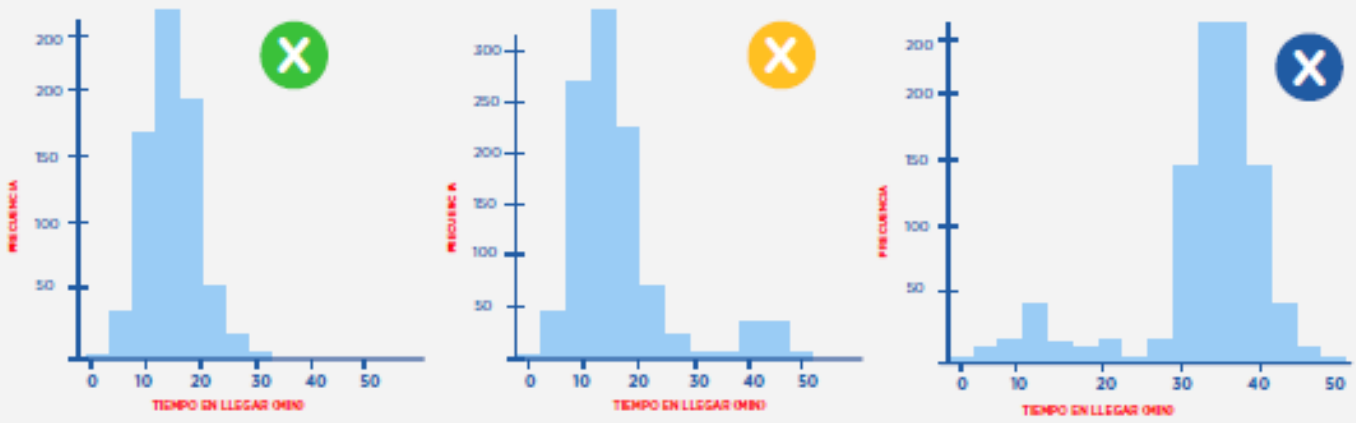


Pero también nos podemos cruzar con datos más complejos.

Por ejemplo, el SAME atiende un gran porcentaje de las emergencias médicas de GBA y CABA. Llegar a tiempo es fundamental para salvar vidas. ¿ Dónde debería estar la base operativa de cada ambulancia de manera de llegar lo antes posible a cada domicilio?

Para responder esta pregunta, hay que tener en cuenta: el tráfico, la geolocalización de las emergencias, el número de ambulancias disponibles, etc. También hay que establecer qué significa llegar lo antes posible a cada domicilio. Ahondemos sobre este último punto. Para ejemplificar la idea, supongamos que el SAME cuenta solamente con una única ambulancia completamente equipada para cierto tipo de lesión (como ser quemaduras graves). Supongamos además que contamos con la dirección donde ocurrieron los últimos 200 accidentes. Si nuevamente los representamos en el mapa vemos que el 90 % de estos ocurren cerca de la cruz verde y el 10 % restante, cerca de la cruz azul.

Ahora, ¿dónde debería estar la base operativa de la ambulancia? ¿Cerca de la cruz verde, cerca del centro indicado por la cruz amarilla o cerca de la cruz azul? ¿Queremos llegar muy rápido a muchos y muy lento a pocos, o queremos ser justos y no discriminar por domicilio dando las mismas oportunidades a todos? ¿Qué políticas desincentivan el crecimiento de las grandes urbes?



ELECTRICIDAD

Uno de los grandes desafíos que tiene la Argentina hoy de cara al futuro es, sin duda, el desarrollo del sector energético. Sabemos que cada día se necesita más energía para que nuestras ciudades funcionen. Con el correr de los años, el aumento demográfico focalizado en grandes urbes, el aumento de la producción industrial, entre otros factores, hacen que la demanda energética aumente sostenidamente. Por otra parte, la relación entre producción y consumo de energía en muchos casos acarrea problemas ambientales vinculados a la ecología y al calentamiento global.

De aquí surgen muchas preguntas que nos podemos hacer:

•¿Cómo se hace para satisfacer la demanda de energía, especialmente en los picos de consumo?

Si bien los avances tecnológicos en el mediano y largo plazo van a jugar un rol fundamental en la creación de energías verdes, aparatos eléctricos, iluminación, vehículos y demás dispositivos que consuman menos energía, hoy en día es necesario proponer políticas públicas que creen incentivos eficientes (para los consumidores y para los productores de energía) para satisfacer la demanda. El Estado cuenta con mucha información relativa a estos problemas que puede ayudar a describirlos, visualizarlos y resolverlos.

Por el momento, concentrémonos en el problema de demanda de energía. Consideremos algunas variables relevantes para entender el consumo de energía.

- **Variables generales:** la temperatura, el día (hábil, feriado o domingo, sábado), el momento del día, la estación del año.

- **Variables particulares**: la localización geográfica de cada cliente, zona (residencia, rural, industrial), densidad poblacional, la tarifa de facturación, el consumo en períodos anteriores, demandas máximas de potencia, curvas de demanda de potencia por cliente.

El objetivo es poder satisfacer la demanda de energía en todo momento. Para ello, la cantidad de energía generada tiene que ser suficiente. Entender las características del consumo permitirá saber cuánta energía es necesaria para cada región.

A continuación, explicaremos cómo dos familias de procedimientos para analizar datos nos pueden ser de utilidad a la hora de entender estos problemas.

EL PROBLEMA DE REGRESIÓN

El problema que consideramos a continuación es poder estimar y/o predecir la demanda energética de un cliente en un determinado momento, teniendo en cuenta las características anteriormente descriptas. Esto puede ser útil para poder predecir si va a haber apagones, dónde es posible que ocurran, si es necesario recurrir a fuentes secundarias de energía, si conviene tener una política que favorezca el uso de la energía en determinados momentos o penalizar su uso excesivo en otros momentos.

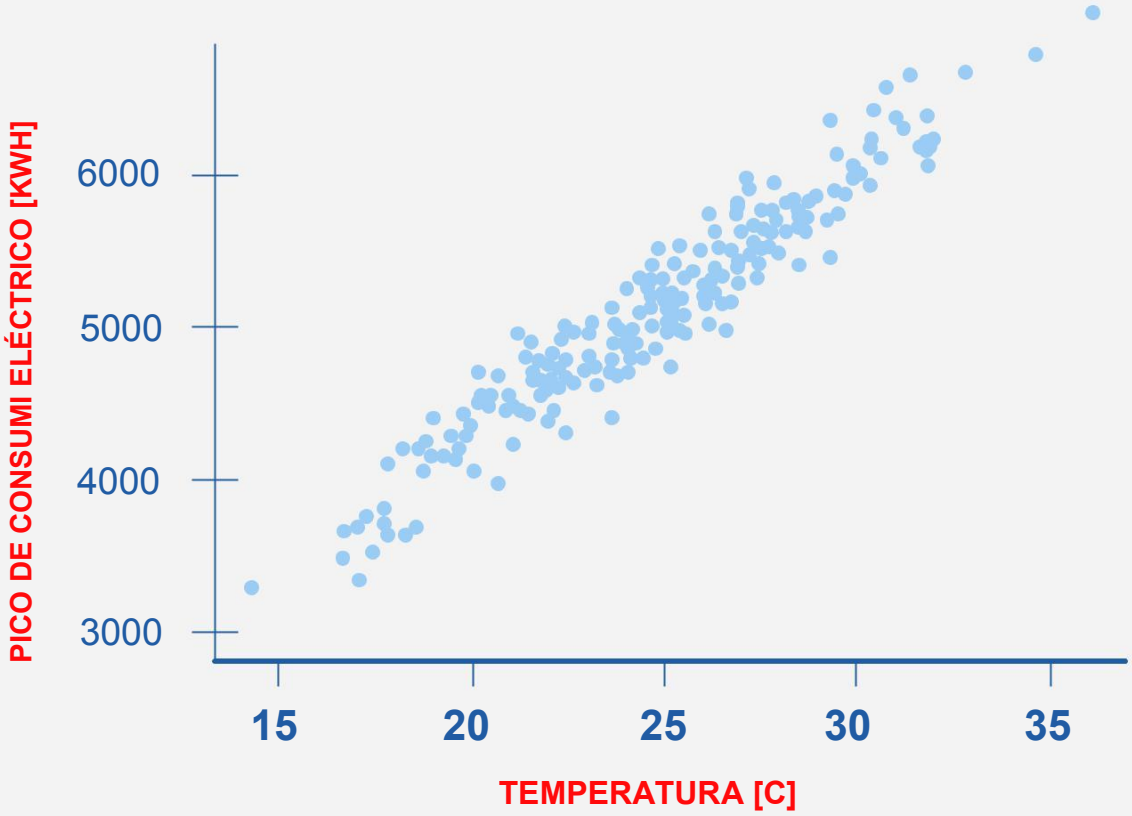
Todos sabemos que una plancha, un lavarropas y un aire acondicionado consumen mucho en relación con el consumo de una lámpara led. El horario de uso de alguno de estos artefactos es flexible (por ejemplo, podemos planchar en cualquier horario del día), mientras que otros no, como las luces que se prenden cuando baja la iluminación natural. A lo largo del día, cada usuario va consumiendo una diferente cantidad de energía eléctrica, y las distribuidoras eléctricas van monitoreando esto e intentan despachar la energía adecuada. El problema surge cuando, en un determinado momento y en forma simultánea, la mayoría de los usuarios quieren consumir mucha energía eléctrica. En ese instante, el pico de demanda en la llamada curva de carga es tan alto que no se puede cubrir y genera así un apagón.

La curva promedio de consumo de energía eléctrica tiene cierto patrón que depende de muchos factores, como el tipo de cliente, la temperatura del día y si es un día laboral o fin de semana. Lo que más interesa de esta curva es justamente el valor del pico de consumo y en qué horario del día se da.

Si bien es claro que todas las variables mencionadas parecen ser relevantes para determinar cuál será la demanda energética en cada momento —y con especial énfasis en el pico de consumo—, también es claro que no todas serán igualmente

importantes. La temperatura es una variable que tiene fuerte incidencia en el consumo eléctrico, principalmente por el uso del aire acondicionado y suele haber más cortes de luz durante el verano en los días de mucho calor que en el resto del año.

Supongamos que contamos con registros de las temperaturas máximas y la máxima demanda energética para cada día del último verano para un usuario.



Del gráfico se desprende que los máximos registros de consumo se dan se dan los días de mucho calor. Por sobre todas las cosas, vemos que hay una relación entre la temperatura y el máximo consumo. Y nosotros queremos aprender de los datos, construir un modelo estadístico que nos permita predecir un resultado o output (en nuestro caso, sería el consumo) sobre la base de datos (más) fácilmente observables de la realidad o input (en nuestro caso simplificado, la

temperatura máxima, pero en el caso original también podríamos haber tenido en cuenta las demás características de los usuarios, climáticas, geográficas, etc.).

Observando nuevamente el gráfico podemos ver que los datos que tenemos se pueden modelar razonablemente bien mediante una recta. A este modelo se lo denomina modelo de regresión lineal simple.

Pensemos que queremos predecir la máxima demanda energética (llamémosla D) a partir de la máxima temperatura diaria (llamémosla T):

$$D \approx \beta_0 + \beta_1 T,$$

donde el símbolo se lee como aproximadamente. Es decir, la máxima demanda energética es aproximadamente una función lineal, una recta que depende de la máxima temperatura.

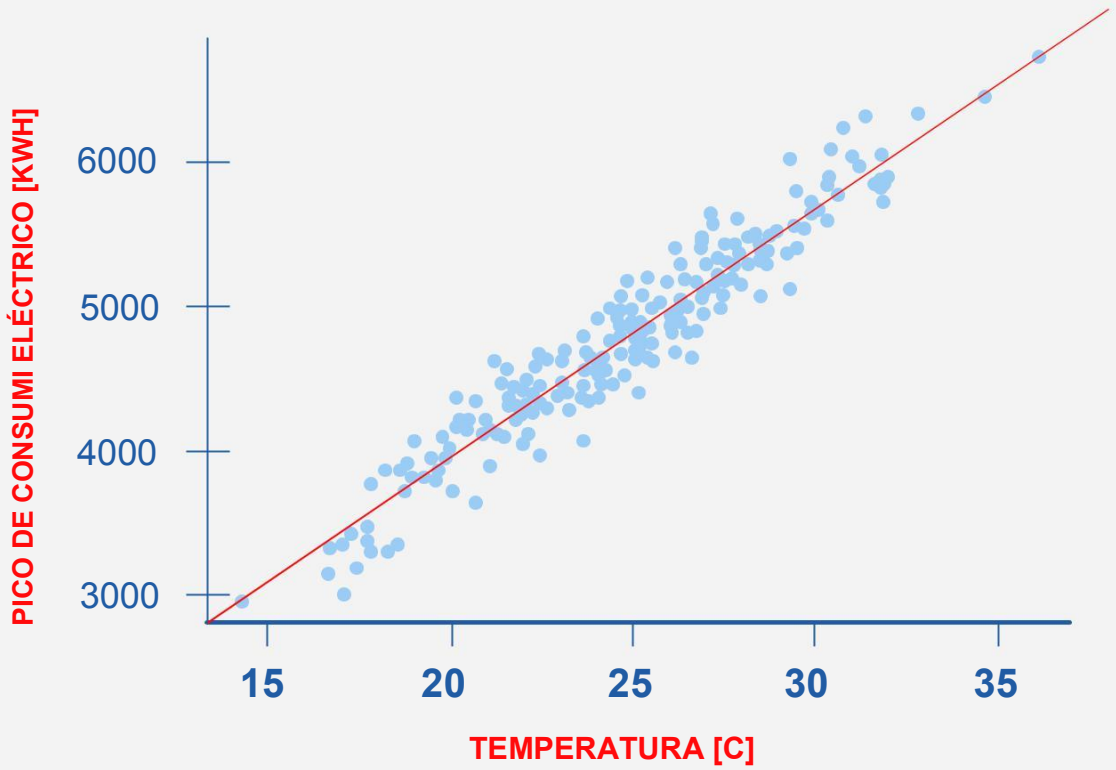
El modelo depende de dos parámetros o coeficientes, y , que debemos estimar a partir de los datos. Luego, si tenemos estas estimaciones y sabemos que mañana la temperatura máxima será de 27 °C, entonces tendríamos una estimación de la máxima demanda energética dada por

$$\hat{D} \approx \hat{\beta}_0 + \hat{\beta}_1 27.$$

Sobre la base de los datos que tenemos graficados, buscamos los mejores valores que puedan tomar los coeficientes del modelo para que que la maxima demanda energetica sea lo más cercana posible a la maxima demanda energética estimada.

Es decir, buscamos los valores de y que hagan que la recta de regresi· n est®globalmente lo más cerca posible de los valores observados: en nuestro caso, la recta que prediga de manera más próxima posible globalmente la m§xima demanda energ®tica.

Hay muchas maneras de medir proximidad, pero por lejos la más difundida se conoce como mínimos cuadrados. Supongamos que tenemos la mejor aproximación a los datos que se puede tener mediante una recta. Entonces, a partir de cada valor que toma la temperatura, podemos predecir la máxima demanda de energía, como se muestra en el siguiente gr§fico:



El parámetro se llama ordenada al origen, que en este caso está estimado por el número 5 (kW) y representa la demanda promedio de energía máxima cuando la temperatura es de 0 °C (una aclaración importante es que, en nuestro caso, únicamente podemos interpretar el modelo para temperaturas entre 15 °C y 35

°C). Es decir, sabemos que hay un piso de 5 kW de demanda de energía.

El coeficiente es la pendiente, que está estimado por el número 180, y mide cuánto aumentará la máxima demanda de energía si la temperatura aumenta en un grado. Es decir, por cada grado que suba la temperatura, la máxima demanda energética aumentará en 180 kWh.

Por tanto, podemos predecir que, si hace 30 °C, la demanda de energía será de 5400 kW.

Alcances de este modelo y márgenes de error

Hasta aquí no hay nada que nos impida hacer esto con cualquier conjunto de datos, pero ¿cómo sabemos que el modelo que estamos ajustando predice bien la realidad?

Es claro que, si tomamos datos de otros veranos y calculamos sus correspondientes rectas de regresión, estas serían ligeramente distintas (aún suponiendo que no hay cambios tecnológicos que hacen que consumamos más o menos energía). Esto se debe a que la estadística busca representar la realidad mediante algún modelo, pero que lógicamente este siempre tiene un error.

¿Qué engloba ese error? El error mide justamente la diferencia que hay entre la realidad y el modelo que propusimos, donde hicimos algunos supuestos:

- El primero fue considerar que las variables utilizadas en el modelo eran las adecuadas para predecir la máxima demanda energética.

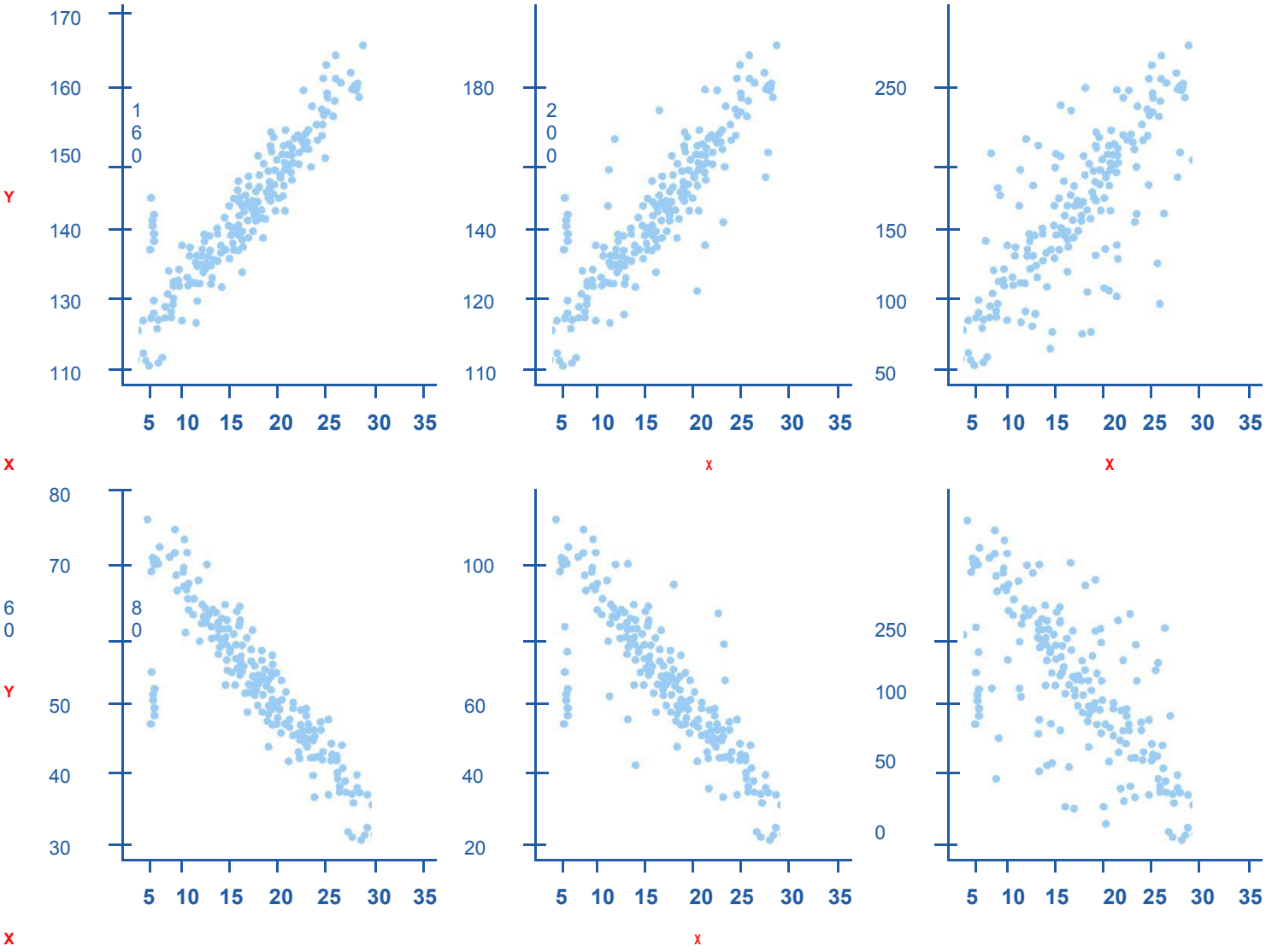
- El segundo fue asumir que el modelo lineal (es decir, ajustar una recta) era adecuado.

El error contiene toda aquella información que es relevante a la hora de determinar la máxima demanda energética y toda aquella información que podría haber sido extraída de las variables utilizando una función que no sea lineal, es decir, una recta. Si ese error es pequeño, entonces podemos estar tranquilos: la información que utilizamos fue adecuada y el modo empleo de la información (es decir, la función lineal) también lo fue.

Algo importante es poder cuantificar el ajuste de un modelo. Típicamente lo que haremos es medir la correlación entre la variable regresora T y la variable de respuesta D. Este coeficiente se conoce con el nombre de R2 y mide la relación lineal entre ambas variables tomando valores entre cero y uno. Si el coeficiente R2 es uno, el ajuste es perfecto; si vale cero, el modelo propuesto no es el adecuado.

En nuestro modelo, el coeficiente R2 es de 0,7. Eso quiere decir que un 70 % de la variabilidad en la máxima demanda energética se puede explicar a partir de la temperatura máxima.

La siguiente figura muestra el valor R2 para distintos ejemplos de datos. Se puede observar que, cuanto más cercano a 1, mejor es explicada la variable Y por la variable X.



Claramente la temperatura máxima es un dato central y relevante a la hora de predecir la máxima demanda energética. Sin embargo, también es deseable poder mejorar esta estimación y esto lo podemos conseguir agregando otras variables, como, por ejemplo:

el tipo de usuario (está claro que un cliente residencial y uno comercial tienen diferente consumo),

el precio al cliente del kWh (a mayor precio menor consumo),
la localización geográfica (el tipo de calefacción y refrigeración depende mucho de la región: en el sur, se utiliza en mayor medida la calefacción central).

De esta manera, nuestra estimación será más precisa.

En general, tenemos una variable que queremos predecir: en nuestro caso, es la máxima demanda de energía. Para ello, contamos con información:

T = máxima temperatura
P = precio del kWh
F = categoría de facturación (residencial, comercial, industrial)
Lat = ubicación geográfica, coordenada de latitud (GPS)
Long = ubicación geográfica, coordenada de longitud (GPS)

A estas variables las denominamos regresoras. Ahora buscaremos predecir la máxima demanda energética, pero utilizando como variables regresoras todas la variables consideradas. Conceptualmente, visualizar este problema es más complicado por tener más de una variable regresora. Sin embargo, las ideas se pueden extender a un contexto general.

Es importante destacar que no tienen todas la misma naturaleza: T, P, Lat, y Long son variables continuas, mientras que F es categórica. Las variables categóricas requieren de un tratamiento especial; por lo tanto, comenzaremos considerando únicamente las variables continuas.

D=+ T+ Lat+ Long+ P+e

Este modelo es mejor que el anterior en el sentido de que nos permitirá estimar la máxima demanda energética de un cliente en determinado momento.

El error , por un lado, engloba todas aquellas variables que inciden en la máxima demanda energética que no utilizamos en nuestro modelo: por ejemplo, aquellas que por simplicidad decidimos omitir u otras variables que no nos imaginamos

que estén incidiendo en la demanda energética, pero que efectivamente lo hagan. Por otro lado, si considerásemos otro tipo de relaciones, el error proveniente de haber impuesto una relación lineal entre las variables se vería modificado.

Si bien en este caso no podemos hacer un gráfico que nos guíe sobre el problema en cuestión, sigue vigente la idea de encontrar los valores de los coeficientes $\beta_0, \beta_1, \beta_2, \beta_3$ y β_4 que minimicen el error de predicción.

Teniendo en cuenta nuestros datos, si tuviéramos los valores de los coeficientes, podríamos en cada caso tener una estimación de D, que llamamos . Luego la diferencia entre la demanda real y su correspondiente estimacióndeben ser próximas, es decir, que nuevamente podemos estimar el error:

Como en el caso anterior, los coeficientes se estiman minimizando la suma de los cuadrados de los errores. Si las estimaciones son buenas, todos los errores serán pequeños, mientras que, si el ajuste es malo, entonces tendremos algunas observaciones con errores grandes.

Sobre la base de nuestros datos, podemos ver que el modelo ajustado es el que sigue:

D+170 T - 4 Lat + 2 Long+3 P

En este modelo, la interpretación de los coeficientes es análoga al caso anterior. Es decir, que la ordenada al origen (intercept) representa el piso de demanda energética, cuando todas las variables son cero; en este caso, es 6 kW.

Luego sabemos que, por cada grado que aumente la temperatura, la demanda aumentará en 170 kW si el resto de las variables permanecen iguales. Por cada

unidad que aumente la latitud, la demanda energética disminuirá en 4 kW. Es decir que, cuanto más al sur estemos, menor demanda habrá (esto podría deberse a características constructivas que hagan que haya una mejor aislación térmica, ventanas más pequeñas, etc. por encontrarse más al sur). Por el contrario, por cada punto que aumente la longitud la demanda aumenta en 2 kW la máxima demanda energética, al oeste la demanda energética es mayor.

Por último, la demanda energética disminuye en 3kW por cada peso que el precio aumenta.

A veces, la interpretación de los coeficientes no es tan simple como hemos expuesto ya que puede haber dependencias entre las variables que hagan que no sea posible una fácil interpretación.

Finalmente, veamos cómo incorporar la variable categórica F (categoría de facturación), que tiene tres niveles: residencial, comercial e industrial. Crearemos variables, que llamaremos dummies: son variables que toman el valor cero en todos los casos, salvo cuando la observación pertenezca a determinada categoría. En este caso, creamos dos variables: FR y FI. La variable FR toma el valor 1 cuando la observación es residencial. La variable FI toma el valor 1 cuando la observación es industrial.

No hace falta tener una variable que indique cuándo es comercial porque será en los casos complementarios. Entonces, el modelo que proponemos es el siguiente:

D=+ T+ Lat+ Long + P+ FR+ FI+e

Nuevamente, los coeficientes se estiman por mínimos cuadrados y tenemos el siguiente modelo:

D5,5+ T+3 Lat-1,5 Long + 4 P+ 5 FR+300 FI

La interpretación de los primeros 5 coeficientes es análoga a la realizada anteriormente. En cuanto a las variables dummies:

Si tenemos una observación cuyo tipo de facturación es residencial, entonces la FR = 1 y FI = 0. En ese caso, sabemos que el máximo consumo base se incrementa en 5 kW.

Si el consumo es industrial, FR = 0 y FI = 1, el consumo base se incrementa en 300 kW.

Por último, si es comercial, FR = 0 y FI = 0, el consumo base no tiene modificaciones.

Este nuevo modelo probablemente tenga un mejor poder de predicción y, por lo tanto, podemos dar por concluido el análisis.

También se podrían eliminar las variables Long y Lat y agregar una variable categórica que sea provincia y modelarla con variables dummies.

A continuación, comentaremos algunas cuestiones sobre las que es bueno reflexionar.

¿Qué variables de entrada son importantes a la hora de predecir la variable de respuesta D? Es decir, ¿qué variables son relevantes a la hora de predecir la demanda de energía?

A menudo, contamos con muchísima variables, que en apariencia pueden estar o no vinculadas al problema. Sin embargo, es probable que únicamente una pequeña fracción de estas sea de utilidad. Luego, es muy importante contar con mecanismos que nos permitan detectar automáticamente cuáles son las variables más informativas en cada problema.

¿Qué tipo de relación tiene cada una de las variables de entrada con la variable que se busca predecir?

Habrá casos en los cuales, cuando la variable de entrada aumenta, la de respuesta también: por ejemplo, a partir de un umbral, la demanda energética aumenta al aumentar la temperatura. Es decir, que tienen una relación positiva. Y habrá otros casos donde la variable de respuesta disminuye cuando la variable de entrada aumenta. Es decir, la relación es opuesta. Por ejemplo, el consumo de gas disminuye al aumentar la temperatura. Esto se debe a que en verano nadie tiene prendida una estufa; los calefones y termotanques requieren menos gas para calentar el agua porque su temperatura inicial es mayor, y la gente es más reticente a hacer comidas que requieran largo tiempo de cocción, entre otras razones. De todos modos, estas relaciones dependen en general de la forma que tenga la relación entre las variables y de la interacción con las otras variables involucradas en el modelo.

En general, buscamos utilizar de la mejor manera posible las variables regresoras para predecir la variable de respuesta: en nuestro caso, la máxima demanda energética. Es decir, que podemos pensar que la demanda es una función de las variables regresoras más un error.

Esto matemáticamente se escribe:

D = f(T, Lat, Long, P, FR,FI) + ε**,**

donde f es una función fija, que generalmente va a ser no lineal de los datos y ε es el error del promedio, que tiene media cero y que es independiente de las variables de entrada del modelo (T, Lat, Long, P, FR,FI). La función f representa un modo de utilizar la información que proveen las variables regresoras, que hasta ahora nosotros asumimos que fue lineal.

Para poder tener buenas predicciones y hacer inferencia de la máxima demanda energética, necesitamos estimar la relación funcional f que existe entre las variables regresoras y que es desconocida. Es decir, necesitamos encontrar el mejor modo posible de utilizar la información provista por las variables regresoras.

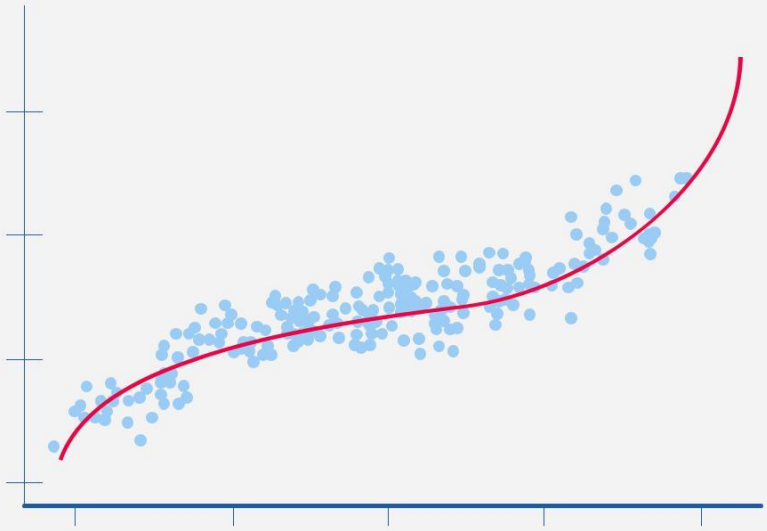
Luego, la pregunta central del aprendizaje estadístico es: ¿cómo estimar la función f?

El enfoque más tradicional es hacerlo mediante un modelo lineal, donde restringimos todas las posibles formas de la relación entre las variables regresoras y le dejamos únicamente tener la forma de un plano. Luego, para resolver el problema, únicamente tenemos que estimar los parámetros que acompañan a las variables regresoras β. Este enfoque, donde se reduce estimar los infinitos puntos que puede tener otra función f a una cantidad finita de parámetros (y poquitos en relación con el tamaño muestral) se conoce como paramétrico.

Estos modelos tienen como ventaja que son sencillos de estimar y de interpretar. Sin embargo, como desventaja, vale mencionar que en ocasiones la rigidez en la forma de la relación entre las variables regresoras y la de respuesta que plantean puede llevar a una mala estimación de esta.

Típicamente, para mejorar el ajuste, se pueden proponer modelos más complejos, que asuman otras formas funcionales. Tales modelos dependerán de más parámetros y, en esos casos, conviene pasar a modelos que den mayor flexibilidad a la forma de relación funcional entre las variables. Estos modelos suelen ser más difíciles de interpretar.

Por otra parte, si se le da mucha flexibilidad al modelo, es decir, se le permite que se adapte demasiado a la forma de los datos, los modelos terminarán teniendo sobreajuste: seguirán los patrones de los errores del modelo de manera muy precisa, perdiendo de este modo el patrón dado por el fenómeno que se está estudiando.



En general, el criterio imperante debe ser mantener el modelo lo más simple posible e introducir complejidad solo en caso de extrema necesidad: por ejemplo, si un modelo más sofisticado (y probablemente más difícil de interpretar) brinda resultados muy superiores, en el sentido de que las predicciones y el modelado general del modelo son mucho más precisos.

El modelo que dominó la literatura en los últimos 200 años fue el modelo de regresión lineal. Esto se debe a varios motivos. El primero es su utilidad: en muchos casos resultó ser muy adecuado para el modelado de diversas situaciones. El segundo es su bajo costo computacional: los cálculos que requiere pueden ser rápidamente hechos a mano o con herramientas muy primitivas de cálculo, eso no lo hace menos atractivo.

Aquí nos encontramos en un punto de inflexión: aprendimos que los métodos más sencillos podrían ajustar peor el modelo (es decir, parecerse menos a la realidad), pero serían sencillos de interpretar, mientras que aquellos más flexibles tendrían mayores dificultades a la hora de la interpretación.

En la práctica, estos balances se resuelven llegando a un compromiso entre ambos. Se buscará tener el método que mejor ajuste tenga, es decir, que nos permita estimar de un modo más preciso la demanda energética, se hace utilizando la menor cantidad de variables posibles para que nos permita entender cuáles son las variables que tienen mayor incidencia en esa demanda energética y nos permitan operar, dentro de los alcances posibles, de modo tal de incidir sobre la demanda. Es decir, no podemos cambiar la temperatura máxima, pero tal vez podamos diseñar estrategias que permitan modificar los hábitos de consumo de los clientes, que nos permitan impulsar medidas para que se beneficie la fabricación o venta de determinados electrodomésticos, automóviles, etc.

Ahora la pregunta natural que sigue es «¿cómo se mide la precisión de un modelo?».

En un mundo ideal, existiría un modelo óptimo capaz de hacer las predicciones más precisas para toda circunstancia. Sin embargo, en la vida real esto no ocurre. En estadística nada es gratis y no existen modelos óptimos. Dependiendo de las características de cada conjunto de datos y problemas, habrá modelos que sean más precisos que otros. Hay modelos que tienen resultados excepcionales para determinados conjuntos de datos y terribles para otros.

Por lo tanto, dado un conjunto de datos y un problema específico, una piedra angular en la resolución de un problema es la elección del modelo óptimo.

Planteamos modelos para representar de modo preciso y sencillo la realidad. Por lo tanto, los errores que cometa el modelo deben ser moderados.

Una aclaración importante: el análisis de regresión no se usa para interpretar las relaciones de causa y efecto entre variables. Sin embargo, este puede indicar cómo se relacionan las variables o en qué medida las variables se asocian entre sí. Al hacerlo, el análisis de regresión tiende a establecer relaciones destacadas que justifican que un investigador con conocimiento lo mire más de cerca.

Veamos el siguiente ejemplo. Una persona podría observar que los días en los cuales se venden más paraguas suele llover; más aún, que la cantidad de paraguas que se venden aumenta cuantos más milímetros de lluvia caen. Sin embargo, es claro que, si dejáramos de vender paraguas, no dejaría de llover. De aquí se desprende que puede ocurrir que la venta de paraguas sea un buen predictor de la lluvia, pero claro está que no es una causa de esta.

En resumen, hemos avanzado sobre las siguientes preguntas:

• ¿Hay alguna relación entre la temperatura máxima, el tipo de facturación, su localización, etc. y los picos de consumo? Nuestro objetivo es ver si los datos que tenemos muestran evidencia de asociación entre estas variables. Si la evidencia fuera débil, entonces no deberíamos tenerlas en cuenta para modelar el consumo.

• ¿Cuáles son las características de los clientes que en mayor medida contribuyen a predecir la demanda estimada? No solo es importante medir el efecto conjunto de todas las características consideradas, sino que además es importante poder establecer cuáles son los efectos individuales de cada una de ellas sobre la variable que se ha de estudiar.

• ¿Con cuánta precisión podemos predecir la máxima demanda energética? Si conocemos cuál será la temperatura máxima para un día en determinado barrio/ciudad y su respectiva conformación de facturación en determinado momento del año, queremos poder predecir un valor puntual para la demanda energética y, además, cotas máximas y mínimas de esta que sean precisas.

• ¿Las relaciones que estamos estudiando tienen una estructura lineal? Si la relación entre la temperatura máxima y la máxima demanda energética es aproximadamente una recta, entonces podremos modelar en forma sencilla la máxima demanda energética. Si no, necesitaremos recurrir a técnicas más complejas para modelar lo que ocurre en diferentes escenarios.

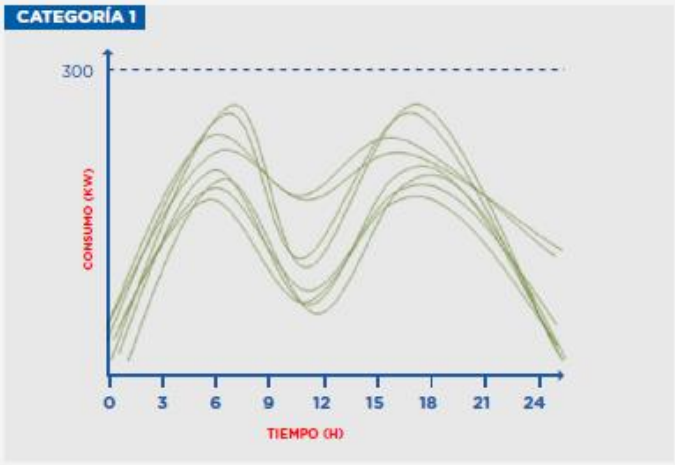
EL PROBLEMA DE CLASIFICACIÓN

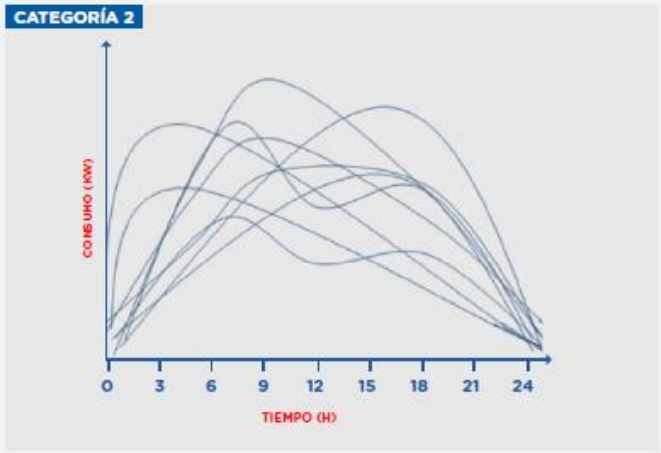
Supongamos que una empresa quiere modificar las categorías actuales de facturación teniendo en cuenta el uso racional de la energía. Con ese objetivo se contrata a expertos en el área y se les pide que clasifiquen a 200 clientes elegidos al azar en k nuevas categorías. Estos expertos toman su decisión teniendo en cuenta características de la curva de consumo, la localización geográfica y variables socioeconómicas del cliente. Los expertos realizaron un procedimiento muy complejo para caracterizar a estos 200 clientes, pero no presentaron en una fórmula (o procedimiento) la manera en que los clasificaron. Para entender el impacto de esta nueva caracterización de facturación, se quiere determinar a qué categoría pertenecería cada uno de los clientes de una dada distribuidora eléctrica.

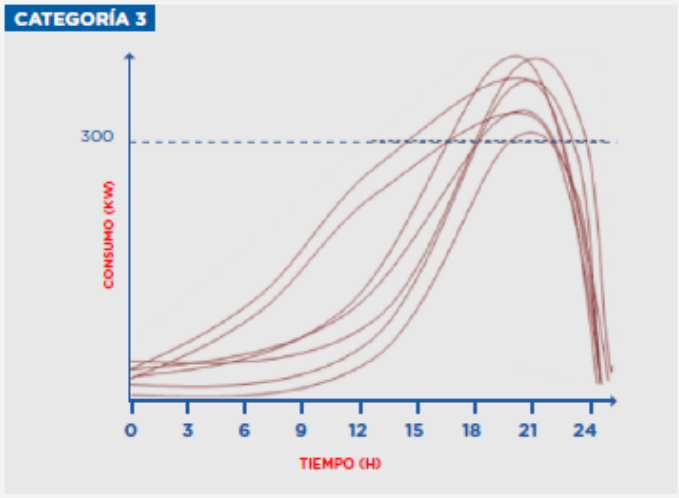
A partir de esta información (200 clientes con nuevas categorías), el objetivo será entonces automatizar la clasificación de absolutamente todos los clientes, es decir, asignarles automáticamente su nuevo estatus de facturación. Si esta tarea se hiciera en forma manual, demandaría mucho tiempo y sería más difícil de lo que parece. Por lo tanto, lo que buscamos es tener procedimientos automáticos para asignar el estatus de facturación a los clientes en base.

Para comenzar, supongamos que los expertos clasificaron a 200 consumidores en 3 tres categorías: C1, C2 y C3. Contaban con las curvas de consumo medio en días hábiles, que brindan información sobre los hábitos de consumo: los picos indican mayores consumos en contraposición a los valles que muestran menores consumos. A partir de estas curvas, se puede también cuantificar la cantidad de energía consumida, es decir, que las curvas medias de consumo son una información muy rica.

Esta muestra sobre la cual se hace la asignación inicial se denomina muestra de entrenamiento.


CATEGORÍA 1
300
CONSUMO (KW)
0 3 6 9 12 15 18 21 24
TIEMPO (H)


CATEGORÍA 2
CONSUMO (KW)
0 3 6 9 12 15 18 21 24
TIEMPO (H)


CATEGORÍA 3
300
CONSUMO (KW)
0 3 6 9 12 15 18 21 24
TIEMPO (H)

El problema que buscamos resolver es el siguiente: tenemos a todos los clientes del sistema eléctrico y buscamos clasificarlos en las tres categorías de facturación que conformaron los expertos: C1, C2 y C3. La técnica más intuitiva para resolverlo y que en la práctica muestra muy buenos resultados se conoce como asignación por vecinos más cercanos.

En este caso, supongamos que tenemos un nuevo consumidor. Conocemos su curva media de consumo en días de hábiles, pero no su categoría de facturación, que es lo que queremos averiguar.

Para ello buscamos en nuestra base de entrenamiento de 200 clientes a los 11 consumidores que sean lo más parecidos posible a él en todas las dimensiones estudiadas. De esos 11 consumidores conoceremos el estatus de facturación. Supongamos que siete de ellos son C1 y cuatro son C2. Entonces, el nuevo usuario recibe siete votos como cliente C1 y cuatro como cliente C2. La asignación se hace por simple mayoría y el nuevo cliente es categorizado como usuario C1. Se lleva a cabo este mismo procedimiento para cada nuevo cliente que se quiera categorizar.

A modo de ejemplo, en la siguiente figura mostramos un caso donde la muestra de entrenamiento es de 11 curvas (4 verdes categoría C1, 3 rojas categoría C2 y 4 azules categoría C3) y queremos clasificar la nueva curva negra utilizando el método de asignación por vecinos más cercanos, pero utilizando solo los 3 vecinos más cercanos.

Las curvas más cercanas a la curva negra son las curvas A, B y C, que son todas categoría C1. Por lo tanto, a este nuevo usuario (curva negra) se le asignaría la categoría C1 ya que todos sus vecinos son de esa categoría.

Luego tenemos caracterizados diferentes tipos de consumidores y además una caracterización sencilla de sus características principales. Sin duda, este es un modelo sobresimplificado, que es muy sencillo de explicar y da rasgos muy generales que describen a los usuarios. Sin embargo, como en cualquier problema de clasificación, podemos clasificar mal a algunos clientes (cometer errores en su clasificación).

Este método es muy popular por su sencillez y su fácil interpretabilidad. Se piensa que la estructura de los árboles de decisión imita el modo en que los seres humanos tomamos decisiones. Su representación gráfica hace que para su aplicación no se requiera ninguna preparación previa. Los árboles pueden lidiar fácilmente tanto con variables cuantitativas como cualitativas. Pero su capacidad predictiva en general no es buena.

Existen técnicas de clasificación basadas en la combinación de árboles binarios que son más precisas: entre ellas, boosting, bagging o random forest, pero están fuera del alcance de este curso.

Como ya aprendimos hasta acá, en estadística nada es gratis y aquello que se gana en mejorar la predicción se pierde en capacidad para interpretar el resultado. Dependiendo del problema en cuestión, la elección tiene que estar basada en el compromiso que asumamos entre habilidad predictiva y la capacidad que tengamos para interpretar los resultados obtenidos.