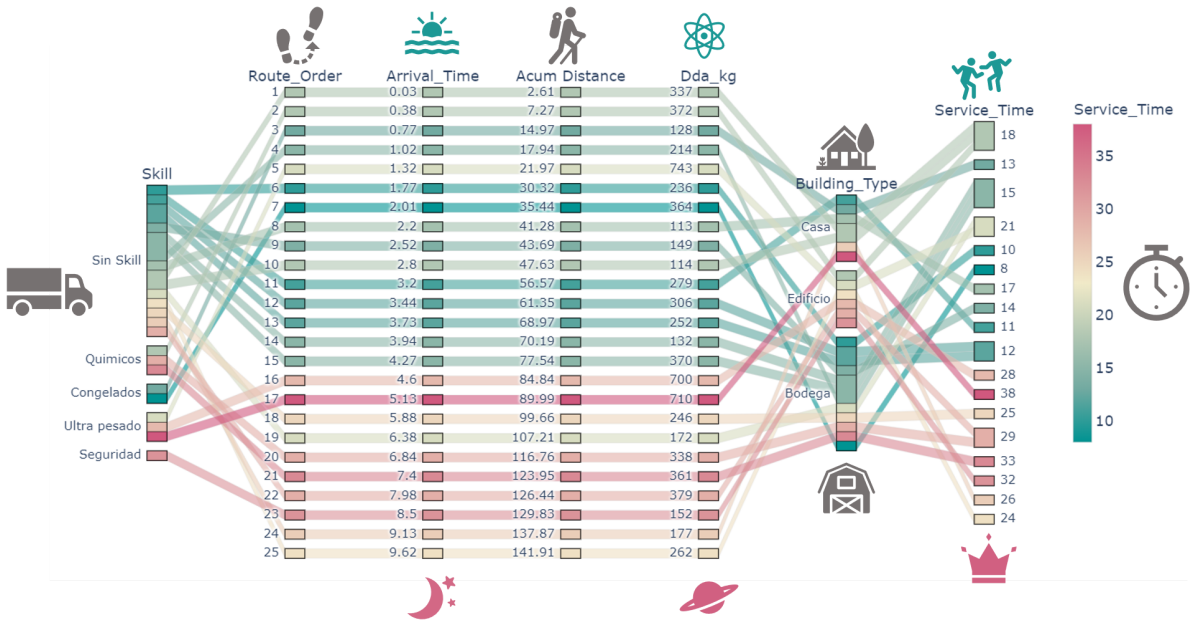
Reduciendo el tiempo de contacto para la entrega de productos en tiempos de Coronavirus con datos de SimpliRoute

Germán Gómez Vargas
Email:gagomezv@gmail.com
23 de mayo de 2020



En tiempos de Coronavirus se hace relevante que el contacto físico entre personas se reduzca a su mínima expresión. Los datos de ruteo de vehiculos proporcionados por SimpliRoute (https://www.simpliroute.com/) permiten extraer información relevante para tomar acciones que reduzcan este tiempo en el cual el virus puede ser transmitido entre quien entrega y quien recibe un producto. Con métodos de visualización de datos y machine learning se encuentra que el tiempo de servicio se va incrementando a medida que se avanza en la ruta de entrega, en general las últimas entregas presentan un mayor tiempo de contacto que las primeras. Así mismo, el peso del producto y la habilidad requerida para la entrega juegan un rol capital en la duración de la visita. El servicio de Aprendizaje Automático o AutoML de Azure Machine Learning (https://azure.microsoft.com/es-es/services/machine-learning/#features) nos provee de una herramienta para explorar una gran variedad de algoritmos, junto con sus parámetros, para poder estimar el tiempo de servicio como función de las características de la ruta y el paquete. Se concluye del análisis realizado que se puede obtener una buena estimación del tiempo de contacto en la entrega, basandose principalmente en el tiempo que se tarda en llegar a la visita, la habilidad requerida para realizarce y el peso del producto. Se recomienda para reducir el tiempo de servicio que se programen las rutas con un máximo de 20 entregas por día y que de ellas, las de mayor peso se hagan al comienzo del recorrido.

DEFINICIÓN DEL CASO

Entre las múltiples dificultades que se presentan durante el **ruteo de vehículos**, un factor importante en la definición de las rutas corresponde al tiempo que necesitará cada vehículo en realizar una entrega de productos en cada visita. Llamamos a este tiempo **"Tiempo de servicio"** y, llevándolo al contexto mundial actual, se puede ver como el tiempo que las personas entran en exposición con otras, dentro de una ruta.

El tiempo de servicio es un dato input para **SimpliRoute (https://www.simpliroute.com/)** , y tiene valor significativo en la formación de rutas. Actualmente, los tiempo de servicios los entregan los mismos clientes que utilizan la plataforma de **SimpliRoute (https://www.simpliroute.com/)**, que por lo general, son tiempos que ellos estiman en base a la experiencia de los choferes o ruteadores. Sin embargo, normalmente estos tiempos no calzan con la realidad, lo cual genera que haya momentos donde las ventanas de tiempo no se cumplen, o peor, donde haya que esperar hasta poder atender una visita en la ruta, incrementando la exposición de quienes realizan los despachos.

Esta situación nos ha motivado a realizar un estudio que nos permita estimar de la mejor manera posible el tiempo de servicio que podría tardar un móvil en cada visita. Para esto, contamos con múltiples registros de tiempos de servicio reales conjuntamente con otras variables de las rutas, que podrían o no, influenciar estos tiempos. Quisiéramos comprobar si es posible analizar estos datos con el fin de obtener información nueva que nos permita mejorar nuestra manera de generar rutas. Favoreciendo así la eficiencia de los negocios, y la seguridad y confort de todas las personas.

CONDICIONES DEL TRABAJO

Con este trabajo, queremos :

1. Analizar la información y determinar si existen características que nos permitan estimar los tiempos de servicio por cliente con el fin de mejorar las rutas.
2. En caso de ser posible, buscar en modelo que realice la estimación de los tiempos de servicio.
3. Evaluar de qué manera el análisis de la información puede impactar en las soluciones.
4. Generar sugerencias que pueden mejorar el negocio.

Los datos a estudiar están entregados en formato .xlsx y corresponden a registros de tiempos de servicio conjuntamente con otra información asociada a las rutas. Se cuenta con 2 tipos de set: Un set Train que servirá para entrenar el modelo, y un set Test que nos permitirá evaluar la calidad de la estimación.

Para realizar esta labor, no hay restricciones en tipo de modelo usado o lenguaje. Todas la asunciones deben ser explicitadas, igualmente que la metodología utilizada.

---

1. Extracción de datos y primer contacto

Primero cada una de las hojas del archivo Excel es transformado en formato .csv y cargado en el servicio **Azure Machine Learning (https://azure.microsoft.com/es-es/services/machine-learning/#features)** para su análisis.

1.1 Diccionario de datos

Tenemos 19 campos tanto en la base de training como de testing, la diferencia es que en el de testing el tiempo de servicio no tiene datos.

Cada registro de cualesquiera de las bases representa un trayecto de una ruta. En la tabla Tabla 1 se presenta el diccionario de datos de las características de cada entrega.

	Field	Description
0	ID_Visit	Unique visit identificator
1	Planned_Day	Day identificator for planned visit
2	ID_Route	Unique route identificator
3	ID_Store	Store identificator
4	ID_Vehicle	Vehicle identificator
5	Route_Order	Indicates the intra route order of the visit
6	Store_X	x coordinate of store position
7	Store_Y	y coordinate of store position
8	ID_Depot	Depot identificator
9	Depot_X	x coordinate of depot position
10	Depot_Y	y coordinate of depot position
11	Dda_kg	Kilograms for deliver in the visit
12	Dda_Vol	Litres for deliver in the visit
13	Skill	Store attribute
14	Building_Type	Store building type
15	Acum Distance	Distance traveled in route
16	Arrival_Time	Hours spend in route
17	Vehicle_Type	Vehicle type
18	Service_Time	Minutes spend in visit

Tabla 1. Diccionario de datos.

1.2 Train dataset

En la Tabla 2 Examinamos los datos de la base Train y observamos que hay 73.858 registros. No detectamos valores faltantes.

	null_sum	null_pct	dtypes	count	mean	median	min	max
Acum Distance	0	0	float64	73858	61.8068	59.815	0.22	182.31
Arrival_Time	0	0	float64	73858	3.37497	3.19	0	11.45
Building_Type	0	0	object	73858	nan	nan	Bodega	Edificio
Dda_Vol	0	0	float64	73858	181.492	140.929	0.0048464	1182.94
Dda_kg	0	0	float64	73858	302.244	267	100	1000
Depot_X	0	0	float64	73858	10336.4	15667	4440	15716
Depot_Y	0	0	float64	73858	9656.14	5226	4480	15227
ID_Depot	0	0	float64	73858	2.43118	2	1	4
ID_Route	0	0	float64	73858	1803.85	1807	1	3600
ID_Store	0	0	float64	73858	1997.35	1991	1	4000
ID_Vehicle	0	0	float64	73858	15.529	16	1	30
ID_Visit	0	0	float64	73858	36929.5	36929.5	1	73858
Planned_Day	0	0	float64	73858	60.6106	61	1	120
Route_Order	0	0	float64	73858	10.961	11	1	25
Service_Time	0	0	float64	73858	18.1526	17	1	44
Skill	0	0	object	73858	nan	nan	Congelados	Ultra pesado
Store_X	0	0	float64	73858	9507.8	9365	3	20000
Store_Y	0	0	float64	73858	10077.2	10011	35	20000
Vehicle_Type	0	0	object	73858	nan	nan	CATNP39A	LIEBX27

Tabla 2. Características de los datos en la base Train. No se detectan valores nulos.

1.3 Distribución de los campos categóricos

Figura 1a

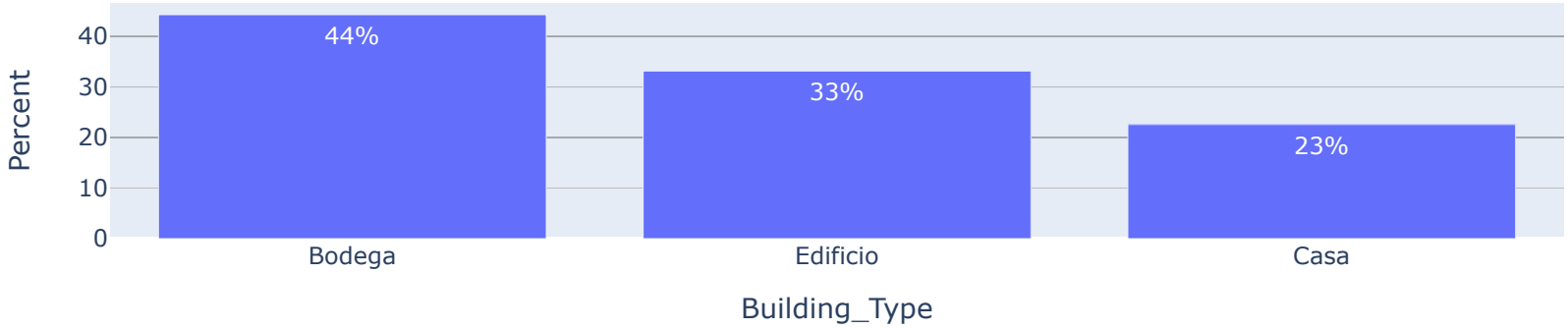


Figura 1b

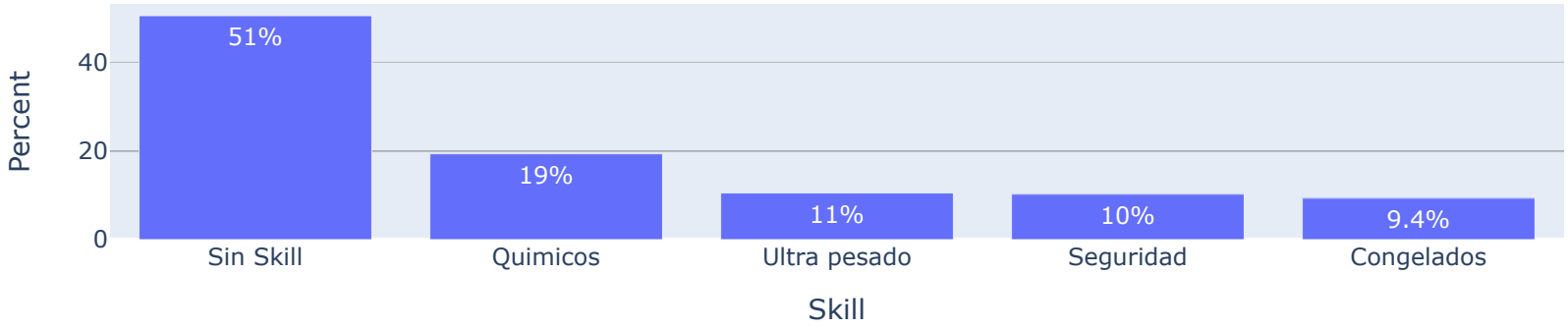


Figura 1c

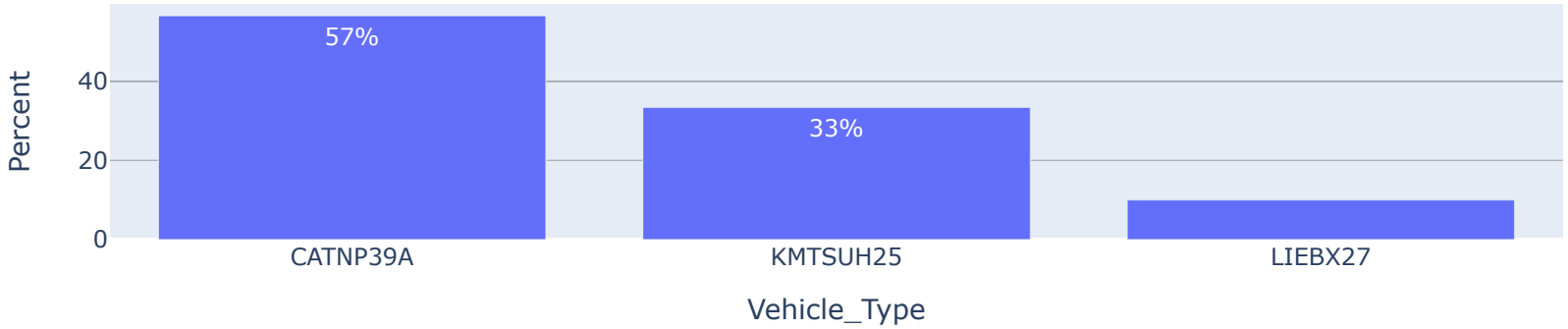


Figura 1a. Distribución de las 73858 entregas en los tres tipos de edificio donde las cuatro depósitos prestan servicio. Aunque principalmente se entrega en bodegas, los otros dos tipos de edificios son significativos.

Figura 1b. Distribución de las 73 858 entregas en los diferentes skills requeridos para realizarlos. Las entregas que no requieren skill conforman la mitad de las entregas que se realizan. Ultra pesado, de seguridad y congelados tienen similar distribución, alrededor del 10%.

Figura 1c. Se utilizan tres tipos de vehículos para hacer las entregas, el principal tipo de vehículo utilizado es el CATNP39A.

Figura 2

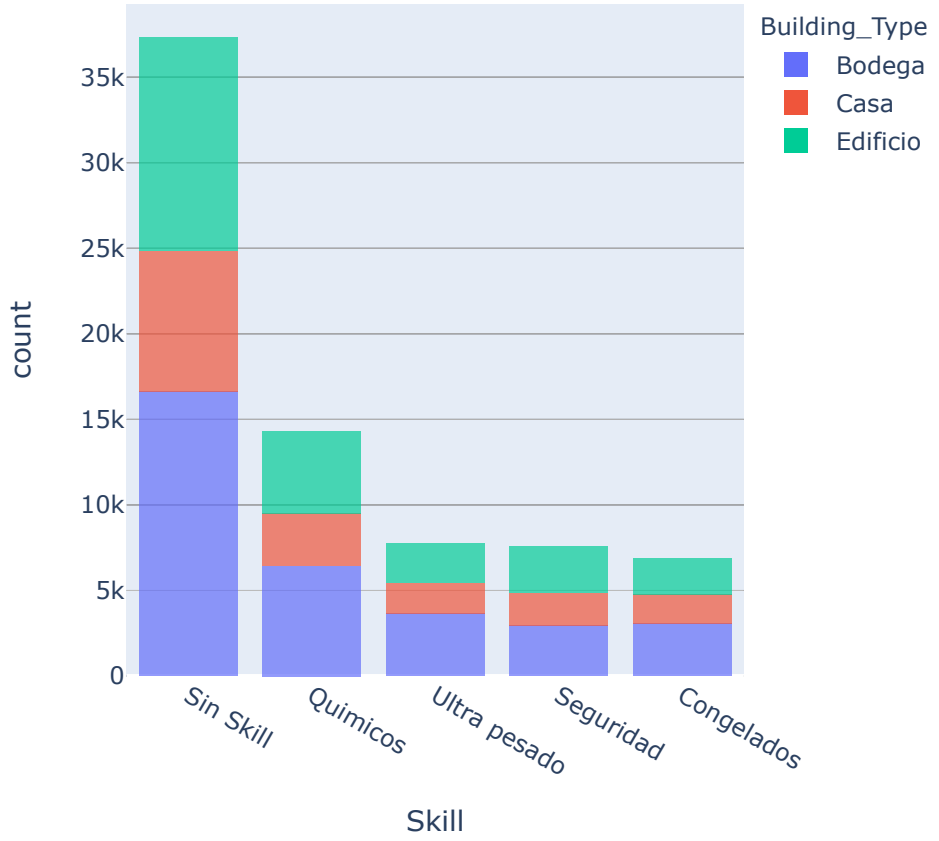


Figura 2. Para profundizar en que tipo de edificio se requieren los diferentes tipos de skill se presenta la distribución de las 73858 entregas por skill y en cada skill la distribución de tipo de edificio. En los tres tipos de edificio se requieren similares skills de entrega.

1.4 Campos numéricos

Graficamos como se distribuyen algunos campos numéricos como función de dos campos categóricos, la skill requerida para la entrega y el típo de edificio en el cual se hace la entrega. También se consideraron los otros campos categóricos pero solo para el típo de edificio se observa una variación en el tiempo de servicio.

Se utilizan diagramas de caja para poder visualizar las distribuciones de las variables numéricas en función de skill y típo de edificio donde se entrega. Los diagramas de caja son una forma estandarizada de mostrar la distribución de datos basada en un resumen de cinco números ("mínimo", primer cuartil (Q1), mediana, tercer cuartil (Q3) y "máximo"). La Figura 3. muestra un ejemplo de una distribución normal representada en un diagrama de caja.

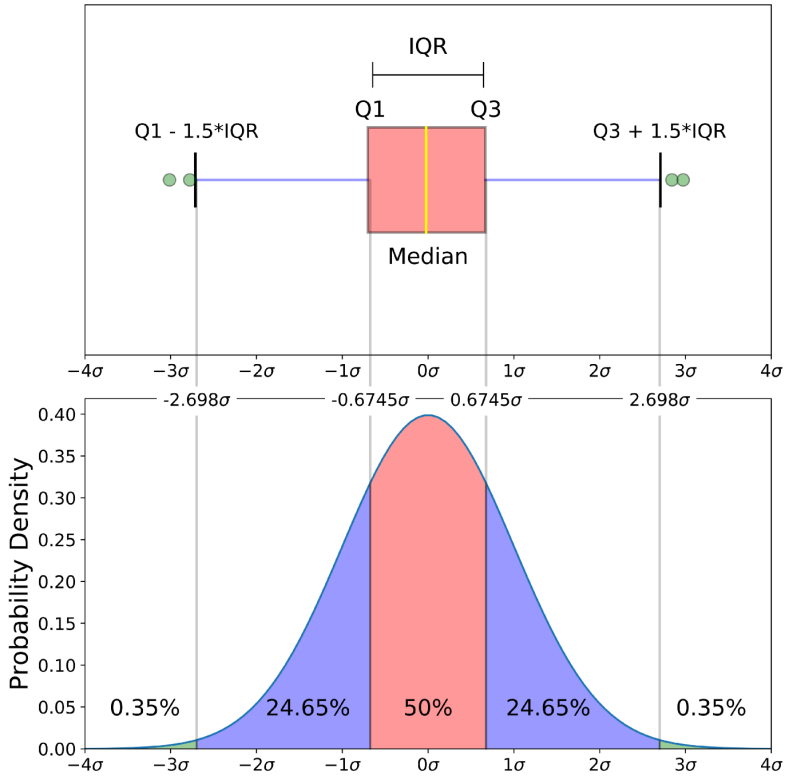


Figura 3. Comparación de un diagrama de caja de una distribución casi normal y una función de densidad de probabilidad (pdf) para una distribución normal

Figura 4a

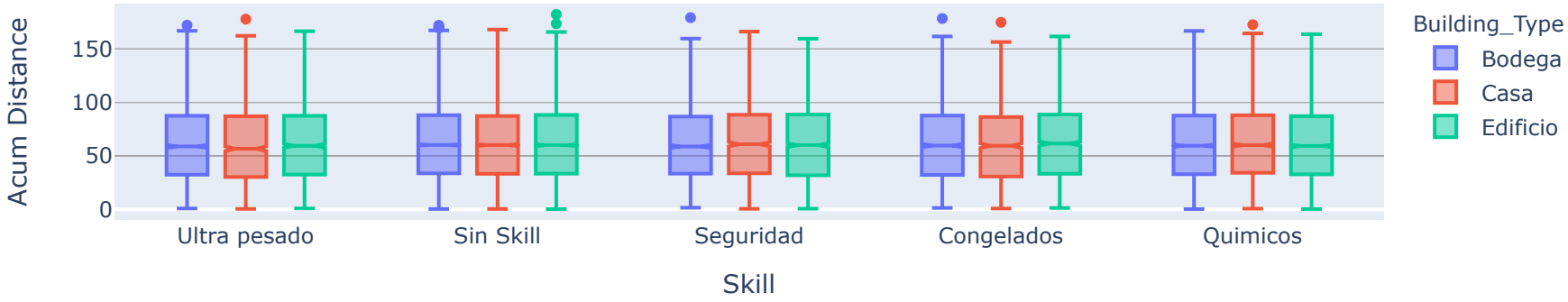


Figura 4b

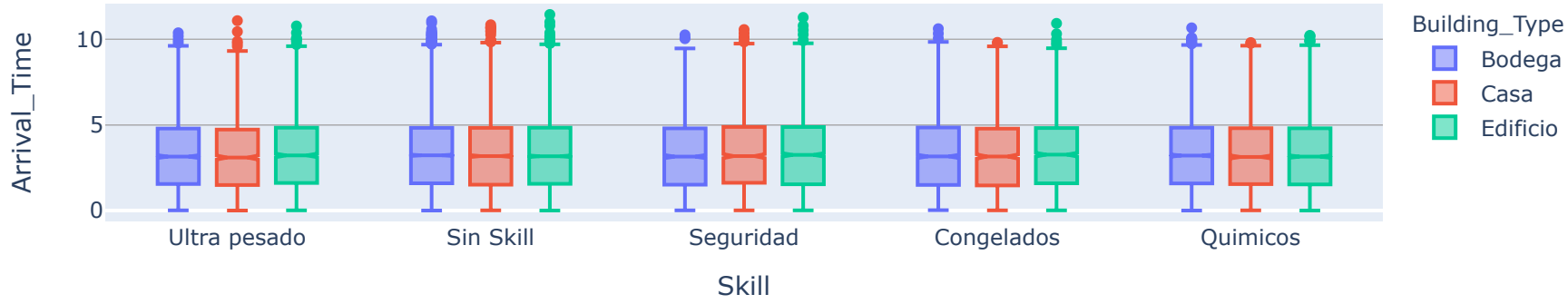


Figura 4c

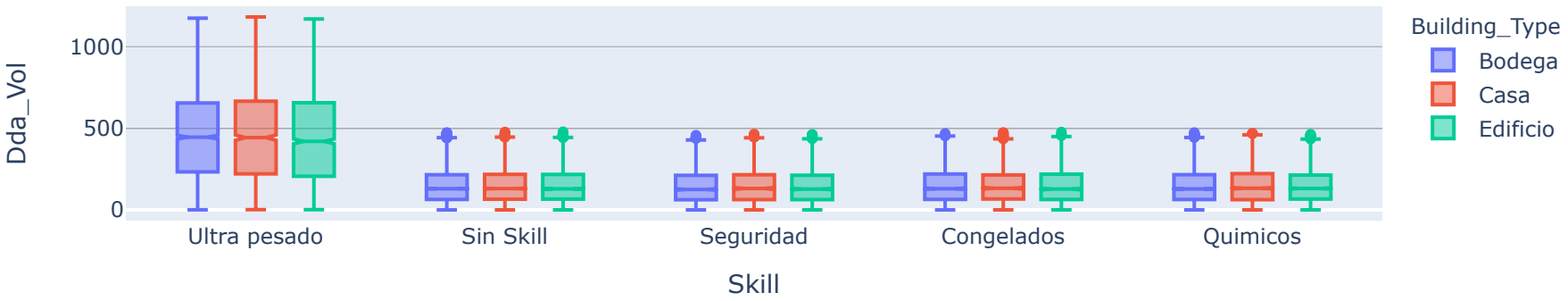


Figura 4d

Figura 4e

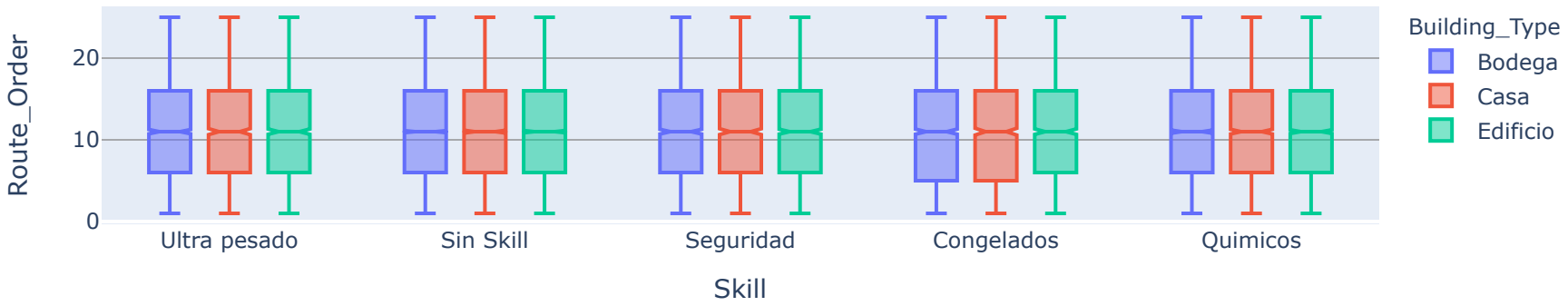


Figura 4f

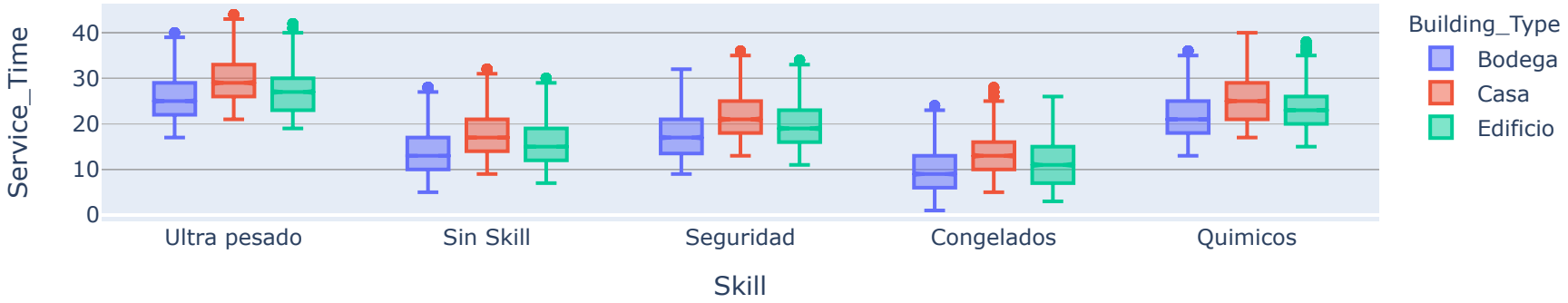


Figura 4a. Diagramas de cajas de la **distancia acumulada** en la ruta en los diferentes skills requeridos para la entrega y separado por tipo de edificio donde se realiza. No se observa una diferencia significativa en la distancia acumulada, con una mediana de 59.8 km, en las diferentes aperturas. La visualización es interactiva y se pueden sacar los tipos de edificio.

Figura 4b. Diagramas de cajas de el **tiempo de arribo** para hacer la entrega en los diferentes skills requeridos para la entrega y separado por tipo de edificio donde se realiza. No se observa una diferencia significativa en el tiempo de arribo, con una mediana de 3.2 horas en las diferentes aperturas. La visualización es interactiva y se pueden sacar los tipos de edificio.

Figura 4c. Diagramas de cajas de el **volumen del paquete** en los diferentes skills requeridos para la entrega y separado por tipo de edificio donde se realiza. No se observa una diferencia significativa en el volumen por tipo de edificio, si separamos por el skill. Pero al separar por skill se hace evidente la diferencia de volumen de los ultra pesados, con una mediana de 435 unidades de volumen, en comparación con las otras skills donde se tiene una mediana de 129 unidades de volumen. La visualización es interactiva y se pueden sacar los tipos de edificio.

Figura 4d. Diagramas de cajas de el **peso del paquete** en kilogramos en los diferentes skills requeridos para la entrega y separado por tipo de edificio donde se realiza. No se observa una diferencia significativa en el peso por tipo de edificio, pero si se observa diferencia en el skill requerido, siendo de mayor peso los de skill ultra pesado con una mediana de 750 kg en comparación con las otras skills donde se tiene una mediana de 249 kg. La visualización es interactiva y se pueden sacar los tipos de edificio.

Figura 4e. Diagramas de cajas de el **orden de la ruta** en los diferentes skills requeridos para la entrega y separado por tipo de edificio donde se realiza. No se observa una diferencia significativa el número de entregas diarias por skill y edificio. La visualización es interactiva y se pueden sacar los tipos de edificio.

Figura 4f. Diagramas de cajas de el **tiempo de servicio** en minutos en los diferentes skills requeridos para la entrega y separado por tipo de edificio donde se realiza. Se observan diferencias significativas en el servicio por skill y entre tipos de edificio. La visualización es interactiva y se pueden sacar los tipos de edificio.

1.5 Correlación

Los coeficientes de correlación se usan en estadísticas para medir qué tan fuerte es una relación entre dos variables. Existen varios tipos de coeficientes de correlación: la correlación de Pearson (también llamada R de Pearson) es un coeficiente de correlación comúnmente utilizado en la regresión lineal.
Es importante recalcar que la correlación no puede diferenciar entre variables dependientes y variables independientes. En otras palabras; la correlación no indica causalidad, solo informa si hay una relación.

La tabla 3 presenta la correlación entre las variables numéricas. Se observan correlaciones significativas entre las variables esperadas, como el tiempo de arribo y la distancia acumulada. Para el tiempo se servicio no se obtienen correlaciones mayores a 0.5 con ninguna otra variable.

Luego de investigar diferentes filtros donde se pudiera obtener mayor correlación con el tiempo de servicio, se encontró que con el skill requerido para la entrega se obtiene el mayor valor. En particular con el skill de seguridad, donde se obtiene una correlación de 0.6, en todos los otros skills la correlación es similar. La tabla 4 muestra estos resultados y en la figura 5 se observa la dispersión entre el tiempo de arribo y el de servicio. La correlación entre estas dos variables se explica porque despues de 5 horas en la ruta no se obtienen tiempos de servicio menores a 19 minutos, pero si superiores a 26 minutos, los cuales no se obtienen en las primeras 5 horas de ruta.

Una correlación útil que aparece es la de el tiempo de arribo a la entrega y el orden de la ruta, porque nos informa de cuánto tiempo se tardan en realizar cierta cantidad de entregas por día. Por ejemplo, el 75% de las entregas número 20 se hacen antes de 7 horas y tienen un tiempo de servicio de máximo 30 minutos

	Acum Distance	Arrival_Time	Dda_Vol	Dda_kg	Route_Order	Service_Time
Acum Distance	1.0	0.95	-0.0057	-0.0075	0.95	0.4
Arrival_Time	0.95	1.0	-0.003	-0.0022	0.98	0.44
Dda_Vol	-0.0057	-0.003	1.0	0.67	-0.0019	0.25
Dda_kg	-0.0075	-0.0022	0.67	1.0	-0.00079	0.37
Route_Order	0.95	0.98	-0.0019	-0.00079	1.0	0.42
Service_Time	0.4	0.44	0.25	0.37	0.42	1.0

Tabla 3. Correlación entre variable numéricas sin ningún filtro.

	Acum Distance	Arrival_Time	Dda_Vol	Dda_kg	Route_Order	Service_Time
Acum Distance	1.0	0.95	0.0028	-0.019	0.95	0.55
Arrival_Time	0.95	1.0	0.0048	-0.016	0.98	0.6
Dda_Vol	0.0028	0.0048	1.0	0.5	0.0018	-0.00048
Dda_kg	-0.019	-0.016	0.5	1.0	-0.015	-0.00044
Route_Order	0.95	0.98	0.0018	-0.015	1.0	0.56
Service_Time	0.55	0.6	-0.00048	-0.00044	0.56	1.0

Tabla 4. Correlación con Skill = Seguridad

Figura 5. Tiempo que se tarda en llegar a hacer la entrega vs tiempo de servicio.

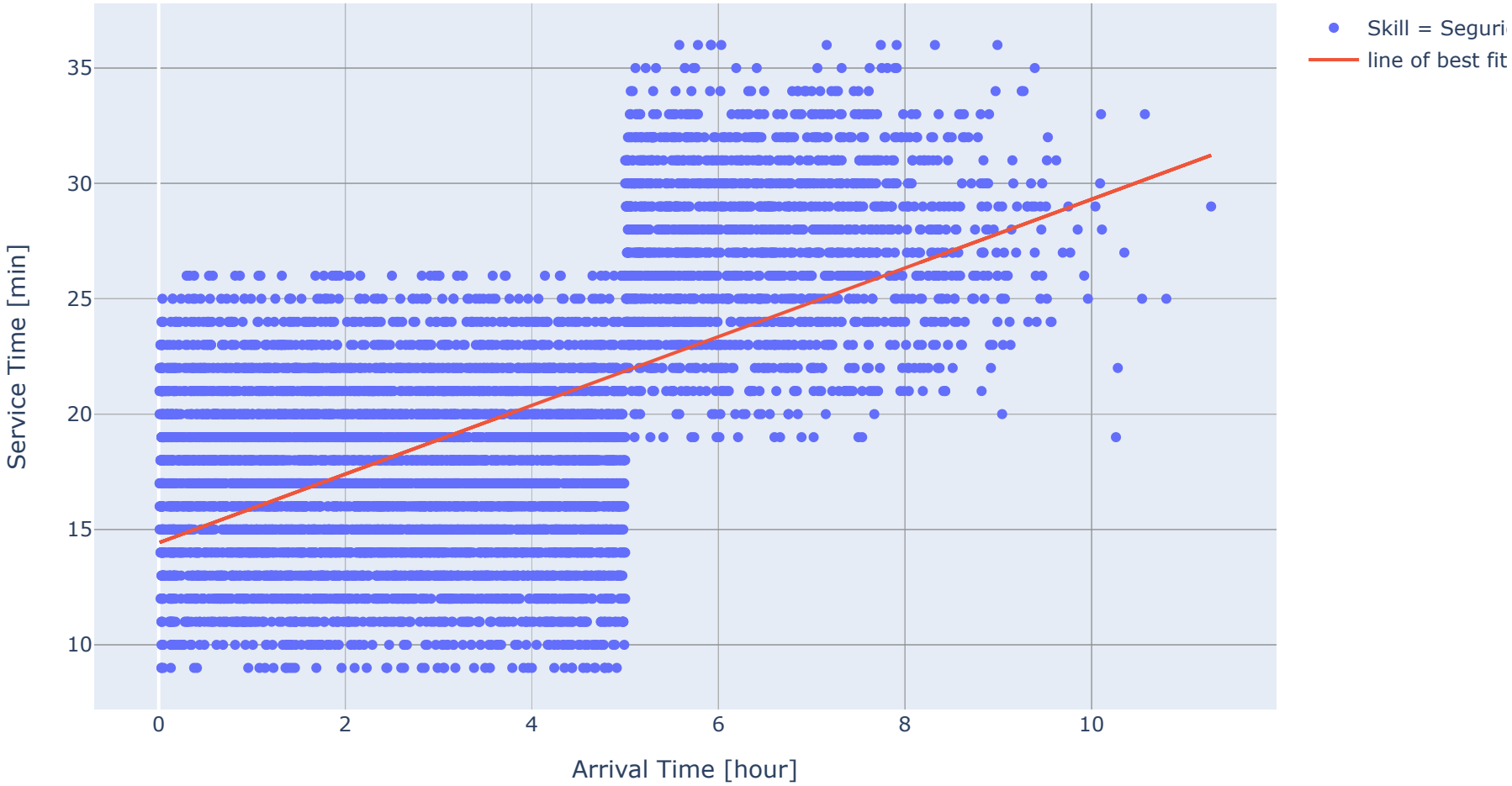


Figura 5. Tiempo de servicio vs. tiempo de arribo para las entregas que requieren el skill de seguridad. Los puntos azules representan cada entrega y la linea roja es un ajuste lineal de los datos que presentan una correlación de 0.6. Esta correlación se da principalmente porque después de 5 horas el tiempo de servicio se incrementa. Se observa el mismo comportamiento en los diferentes skills, pero en seguridad es donde está más acusado.

1.6 Algunas observaciones de la sección de extracción de datos y primer contacto:

- Se planifican entregas en 120 días para entregas desde 4 depósitos a 4000 tiendas. Son 30 vehiculos los que realizan estas entregas.
- En los gráficos de caja no se observa diferencia en el comportamiento de los cuatro depósitos.
- Los vehiculos salen de los depósitos y realizan máximo 25 entregas por día, recorren mínimo 0.22 Unidades de Distancia (UD) y máximo 182.31 UD, pero normalmente solo recorren 60 UD. Cada día máximo hacen la ruta en 11.45 horas.
- Para la mitad de las entregas no requieren skill, pero un 11% son paquetes pesados, los cuales típicamente pesan 750 kg, es decir, 500kg más que en los otros paquetes.
- Los paquetes pesados requieren, típicamente, un tiempo de servicio de 27 minutos, esto es unos 17 minutos más que el tiempo de servicio para un paquete congelado.
- El tiempo de servicio no tiene una correlación significativa con ningún campo numérico cuando no se filtra por skill. La mayor correlación está con el tiempo de arribo, que a su vez está estrechamente correlacionado con la distancia acumulada y la orden de entrega, lo que tiene mucho sentido, a medida que se va realizando la ruta se recorre más distancia y el tiempo va pasando.
- El tiempo de servicio muestra correlación con el tiempo de servicio al filtrar por skill, se observa una correlación pequeña de 0.6 en el skill de seguridad, para los otros skills es similar la correlación.

2. Mapa de las entregas

Para visualizar geográficamente la posición de los depósitos de inicio y sus zonas de entrega graficamos un mapa de contornos donde las curvas de nivel representan las zonas donde más se hacen entregas.

De cada uno de los depósitos 1 y 2 salen 960 rutas, de los restantes 4 y 3 salen 840, ver tabla 5.

En la figura 6 se presenta un mapa que emerge de los datos de forma cuadrada con los cuatro depósitos formando un rectangulo, sus zonas de entrega, aunque se solapan levemente, están visiblemente demarcadas. De el depósito 4 (violeta) se hacen entregas en zonas más localizadas, mientras de el depósito 3 (verde) se reparte a puntos más uniformes.

	Porcentaje de despachos por depósito
2.0	26.7%
1.0	26.7%
4.0	23.3%
3.0	23.3%

Tabla 5. Distribución de los 3600 despachos por deposito.

Figura 6. Mapa de contornos de las 73858 entregas

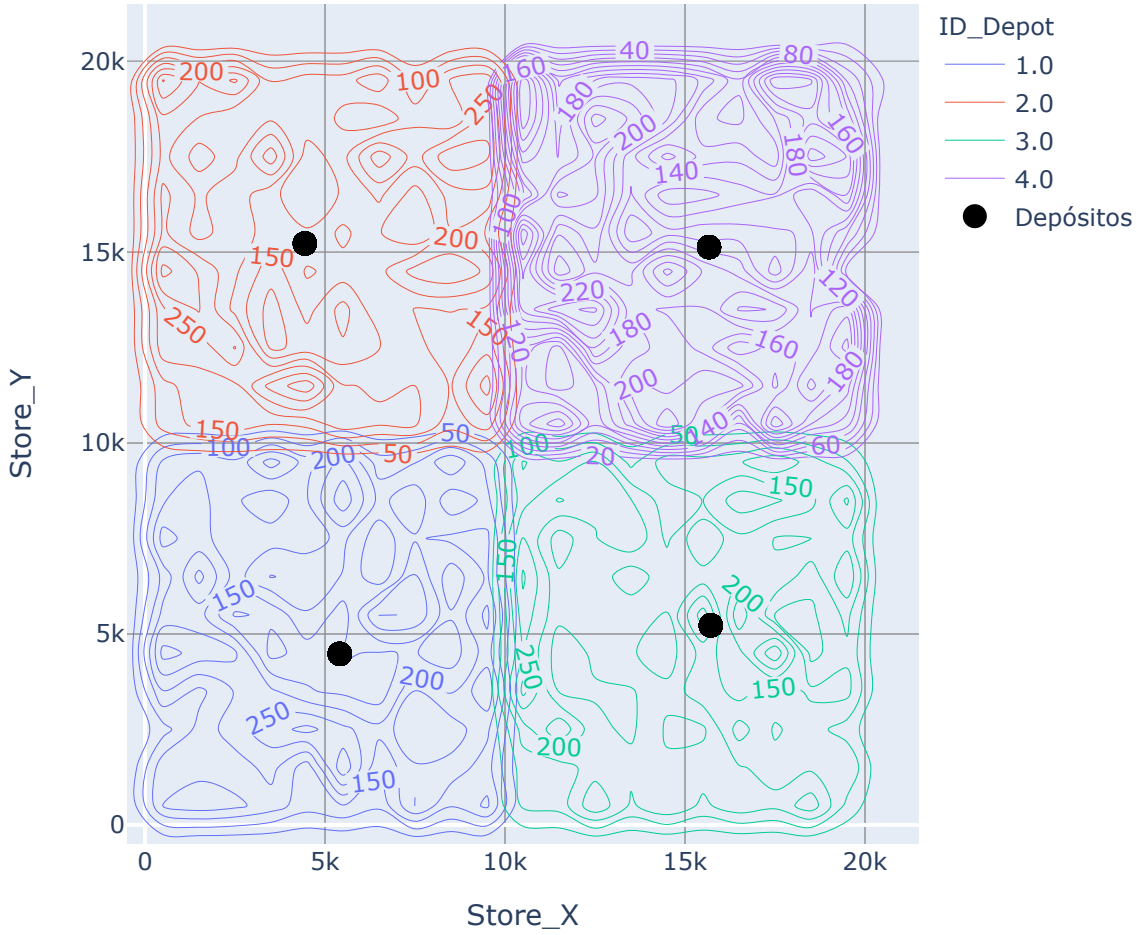


Figura 6 Mapa de contornos donde las curvas de nivel representan las zonas donde más se hacen entregas. Existen cuatro zonas demarcadas que se diferencian el mapa por color, los puntos negros indican la posición de los depósitos desde donde parten las rutas.

3. Visualización de rutas

Para visualizar las rutas se intentaron diferentes métodos. A continuación se muestran dos que ayudaron a dar mayor información sobre la inferencia del tiempo de servicio.

3.1 Coordenadas Paralelas

La definición de Wikipedia:

"Las coordenadas paralelas son una forma común de visualizar geometría de alta dimensión y analizar datos multivariados.
Para mostrar un conjunto de puntos en un espacio n-dimensional, se dibuja un fondo que consta de n líneas paralelas, típicamente verticales e igualmente espaciadas. Un punto en el espacio n-dimensional se representa como una polilínea con vértices en los ejes paralelos; La posición del vértice en el eje i-ésimo corresponde a la coordenada i-ésima del punto.
Esta visualización está estrechamente relacionada con la visualización de series de tiempo, excepto que se aplica a datos donde los ejes no corresponden a puntos en el tiempo y, por lo tanto, no tienen un orden natural. Por lo tanto, diferentes disposiciones de eje pueden ser de interés."
https://en.wikipedia.org/wiki/Parallel_coordinates (https://en.wikipedia.org/wiki/Parallel_coordinates)

Metodología
Seleccionamos de forma aleatorea un día y un vehiculo para visualizar su ruta del día y así podemos crear hipótesis del comportamiento de las rutas. Luego vemos lo que hicieron 10 vehiculos ese mismo día, para observar patrones con mayor estadística y poder corroborar, o no, las hipótesis hechas con una sola ruta.

3.1.1 Parallel category

La figura 7a Es una visualización de una ruta donde se escogen un vehículo y un día de entrega de forma aleatoria. Se observa que en una ruta las entregas pueden requerir de todos los skills. La parte de abajo de la figura 7a tiende a ser más roja (mayor tiempo de servicio) que la superior, la cual tiende a ser verde (menor tiempo de servicio). Esto coincide con el orden de entrega, las primeras en la parte superior y las últimas al final del día. También se observa en la misma figura que en una ruta se entrega a diferentes tipos de edificios. Estas **hipótesis** se analizan en la figura 7b donde se visualizan las rutas hechas el mismo día por diez vehículos más. Con el puntero sobre la coordenada *Route_Order* de la figura 7b se apunta hacia segmentos partículares de tiempo de servicio. Se confirma la hipótesis de que al final de la ruta se tiene mayor tiempo de servicio. Así mismo se puede observar en la figura 7b que el tipo de edificio donde se hace la entrega es indistinto del orden de entrega o el skill.

Figura 7a. Coordenadas categóricas paralelas.
Ruta individual de un vehiculo aleatorio:11 un día aleatorio:38

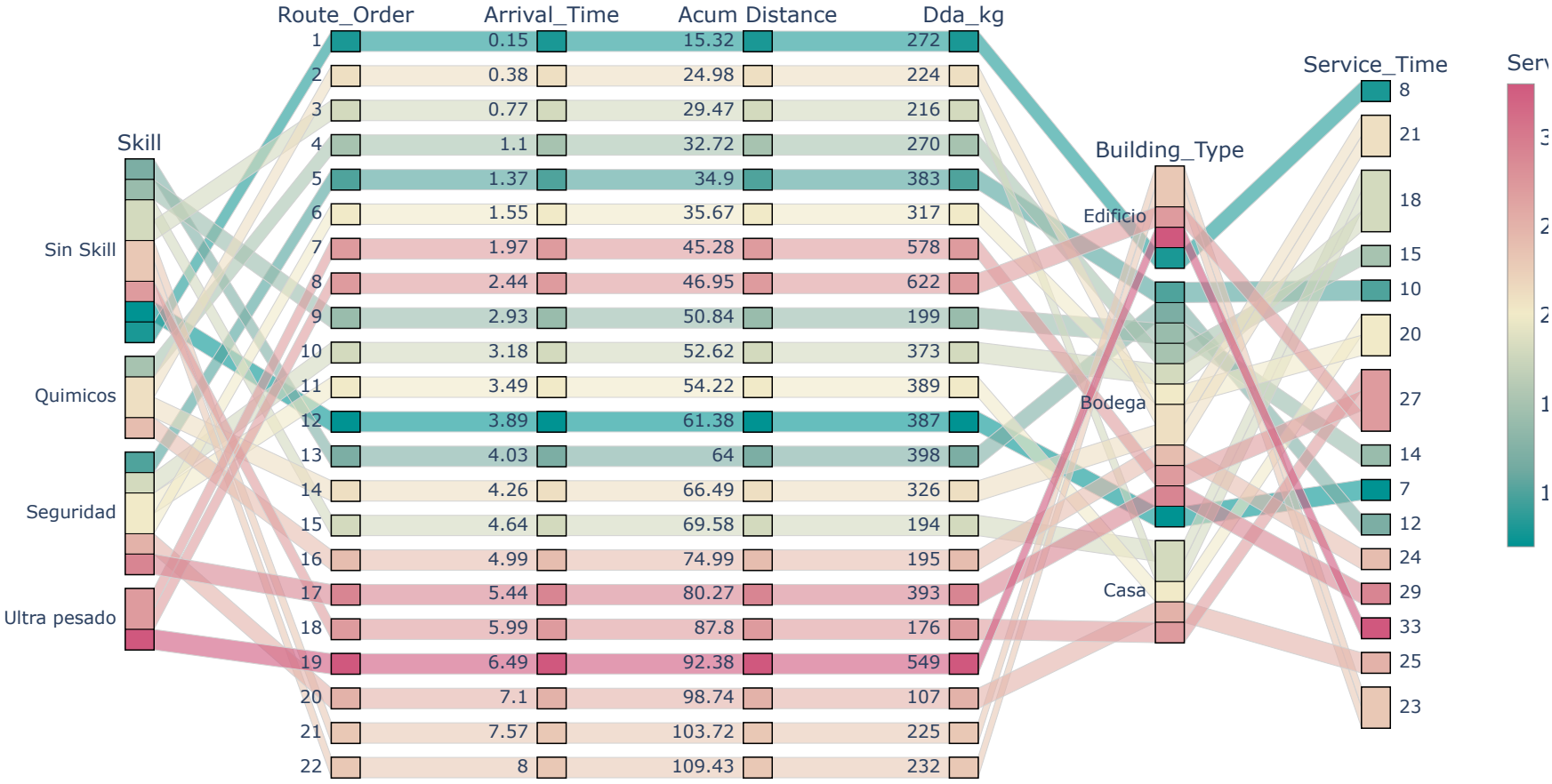
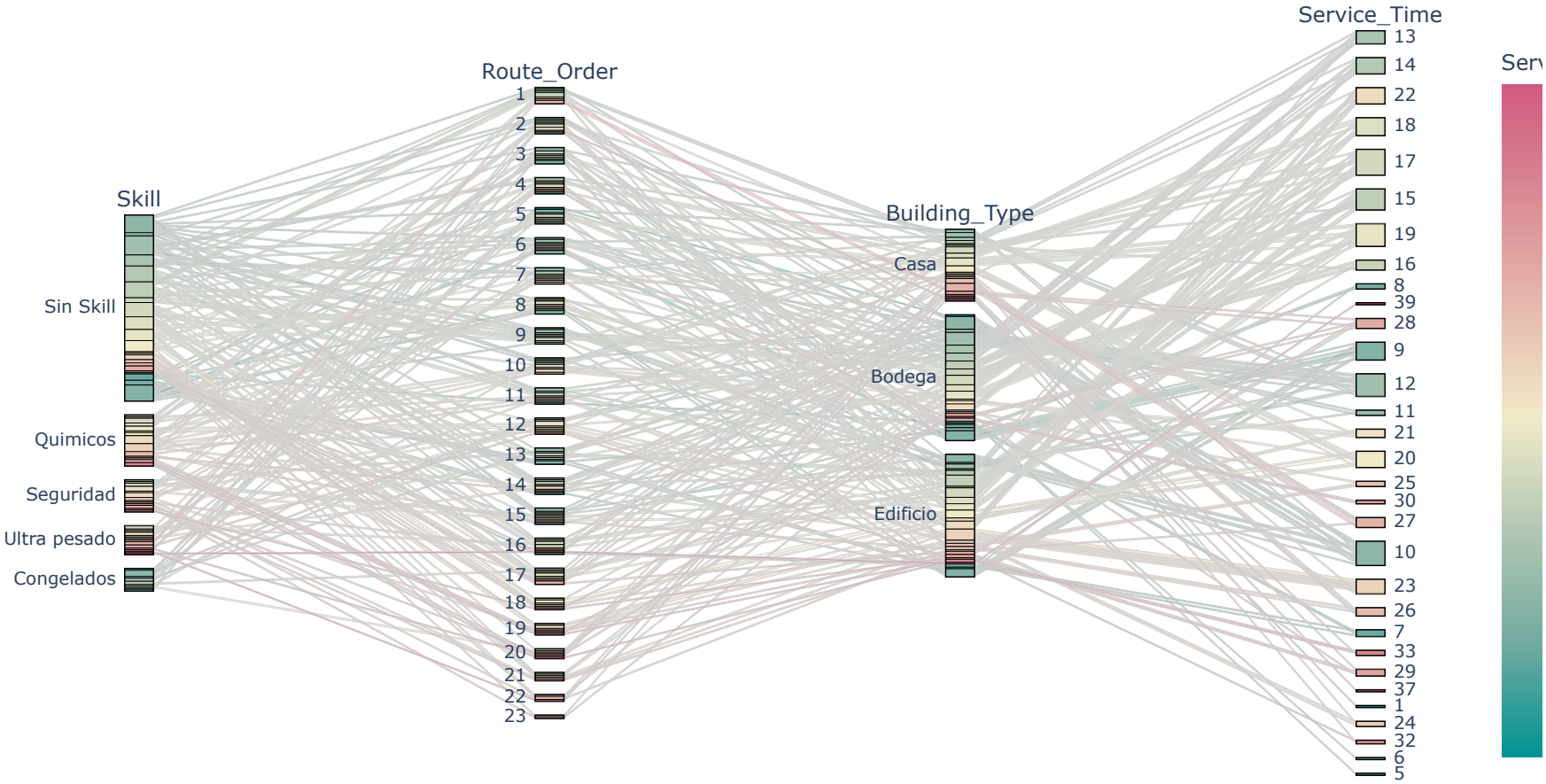


Figura 7b. Coordenadas categóricas paralelas.
Rutas de diez vehiculos el mismo día de la figura 7a



3.1.2 Parallel numerical

En la figura 8a el orden en cada coordenada lo determina el valor numérico de la misma. En esta visualización de una ruta se muestra toma el mismo vehiculo y día de la figura 7a. Se observa que a medida que se avanza en el orden de la ruta, el tiempo de arribo va aumentando, entre las dos primeras coordenadas las lineas se desvían de ser horizontalmente paralelas cuando tienen un mayor tiempo de servicio. La principal hipotesis que se plantea con esta visualización es que los paquetes más pesados son los que mayor tiempo de servicio requieren. Pero cuando se entregan en la mañana generan menor tiempo de servicio. En la figura 8b analizamos esta hipótesis. La figura 8b es la visualización de las rutas de diez vehículos diferentes en el mismo día de entrega de la figura 8a. Para indagar en las hipótesis podemos inspeccionar las diferentes líneas y observar que efectivamente las entregas de mayor peso son las que requieren mayor tiempo de servicio. Las entregas de mayor peso se entregan indistintamente en cualquier parte del orden de ruta, pero efectivamente en la parte de abajo (primeras entregas) las entregas más pesadas tienden a ser amarillas, mientras las de la parte de arriba (últimas entregas de la ruta) son más rojas. Se comprueba que entregar en la mañana los más pesados tiende a reducir el tiempo de servicio.

Figura 8a. Coordenadas numéricas paralelas. Ruta car: 11 day: 38

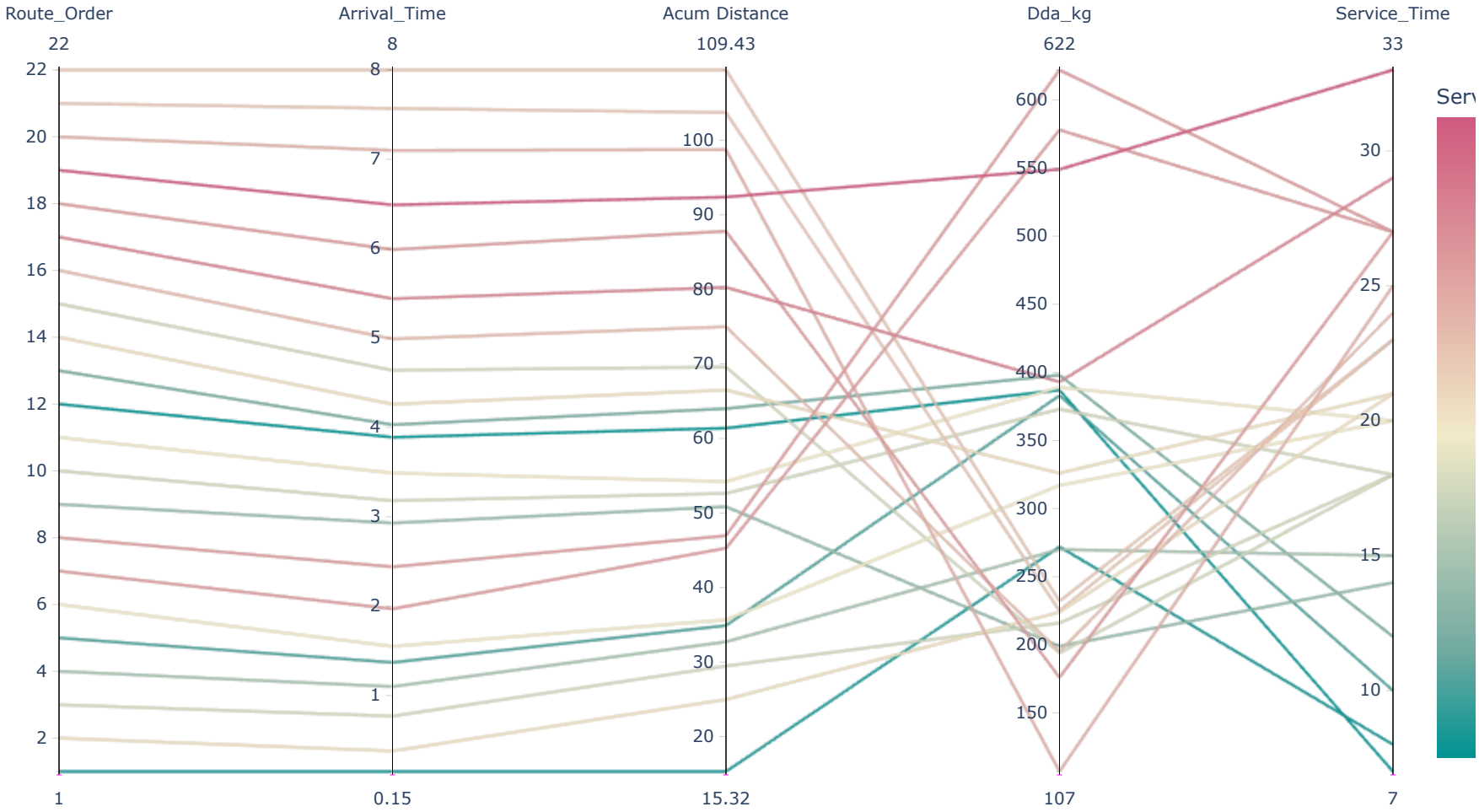
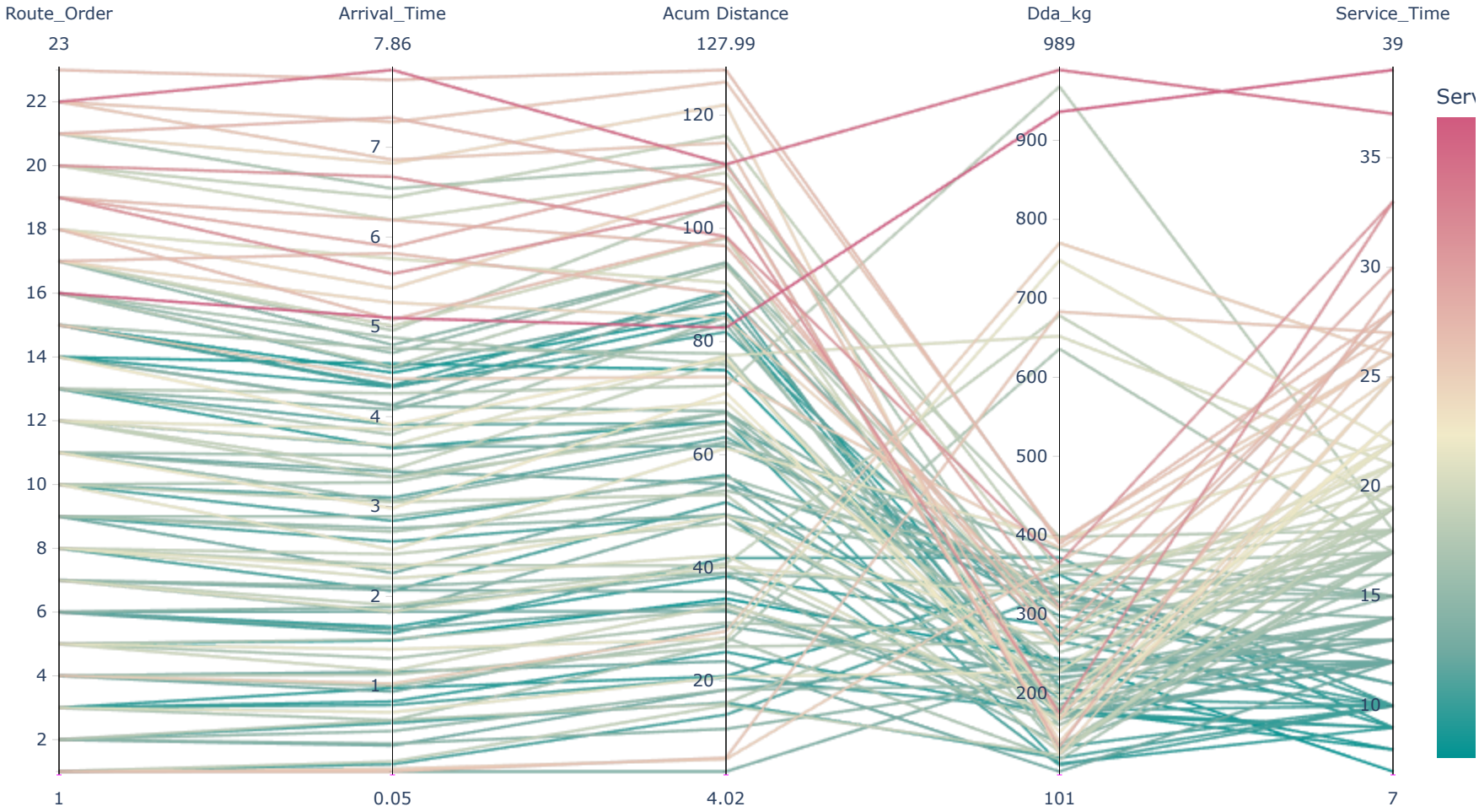


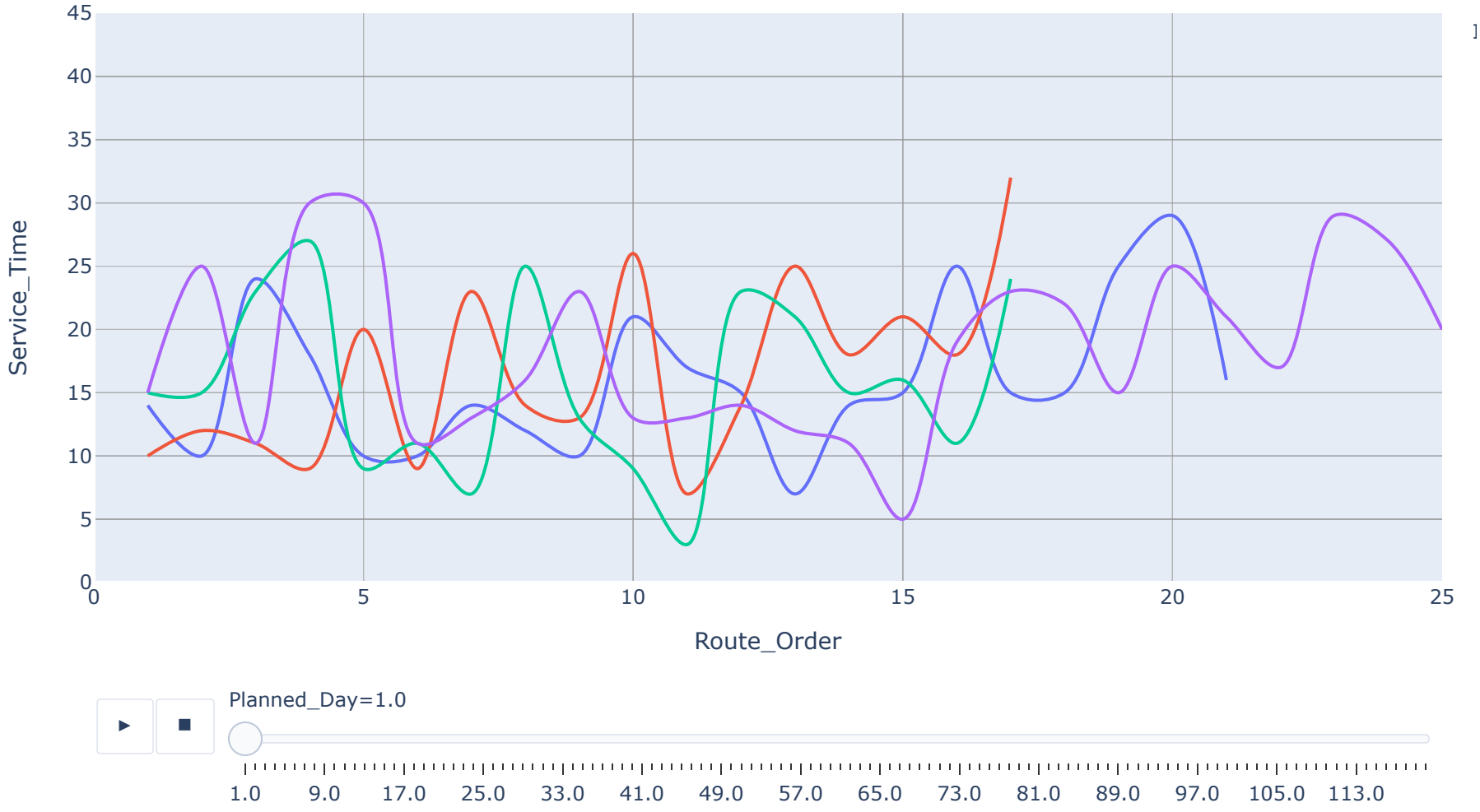
Figura 8b. Coordenadas numéricas paralelas. Rutas de diez vehiculos day: 38



3.2 Visualización colectiva de rutas

Para observar patrones en la ruta se presenta la animación de la figura 9 donde se grafica el tiempo de servicio como función del orden de la ruta y se evoluciona con el día de la entrega. Esta gráfica nos ayuda a visualizar que, luego de 20 entregas el tiempo de servicio es mayor, pero que no todos los días los vehiculos tienen más de 20 entregas programadas. Por simplicidad de la animación se escogieron cuatro vehiculos de forma para representar. Cabe recalcar, como ya se mencionó anteriormente, que el 75% de las entregas número 20 se hacen antes de 7 horas y tienen un tiempo máximo de servicio de 30 minutos. Luego podremos hacer la misma visualización con los datos de test, usando el modelo de predicción de tiempo de servicio. Link a las curvas estimadas de los mismos 4 vehiculos.

Figura 9. Tiempo de servicio como función del orden de la ruta evolucionando con el día.



4. Predicción del tiempo de servicio

Como los datos se observan limpios pasamos a usar AutoML de Azure para encontrar el modelo y sus parametros que mejor predigan el tiempo de servicio. Así mismo, el servicio de AutoML nos ofrece un modulo de explicabilidad donde nos presenta las variables más relevantes que encuentra para hacer la predicción

4.1 AutoML

Azure AutoML (https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml) es un servicio de la nube de Microsoft, quienes lo definen como:

"El aprendizaje automático automatizado, también conocido como ML automatizado o AutoML, es el proceso de automatización de las tareas iterativas que requieren mucho tiempo del desarrollo del modelo de aprendizaje automático. Permite a los científicos de datos, analistas y desarrolladores crear modelos de ML con alta escala, eficiencia y productividad, todo mientras mantiene la calidad del modelo. El ML automatizado se basa en un avance de nuestra división de Microsoft Research."

El paper donde se sustenta el método es **Probabilistic Matrix Factorization for Automated Machine Learning (https://arxiv.org/abs/1705.05355)** Nicolo Fusi, Rishit Sheth, Huseyn Melih Elibol (2018).

La tabla 6 presenta la correlación Spearman de los 14 mejores algorítmos encontrados por AutoML. El servicio permite escoger la métrica que se usa para rankear los algoritmos, se escogió la correlación Spearman pero es equivalente a la de Pearson presentada en la sección 1.5. El mejor modelo encontrado es un *Stack Ensemble* que consiste en considerar weak learners heterogéneos, los que aprenden en paralelo y los combina entrenando un metamodelo para generar una predicción basada en las diferentes predicciones de los modelos débiles, una lectura para entender este típo de algoritmos se encuentra aquí (https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205).

La figura 10 presenta los resultados del análisis de explicabilidad del modelo stack ensamble encontrado por AutoML. El gráfico confirma lo que se venia observando en las visualizaciones de datos con respecto a las características de las entregas que influyen en el tiempo de servicio.

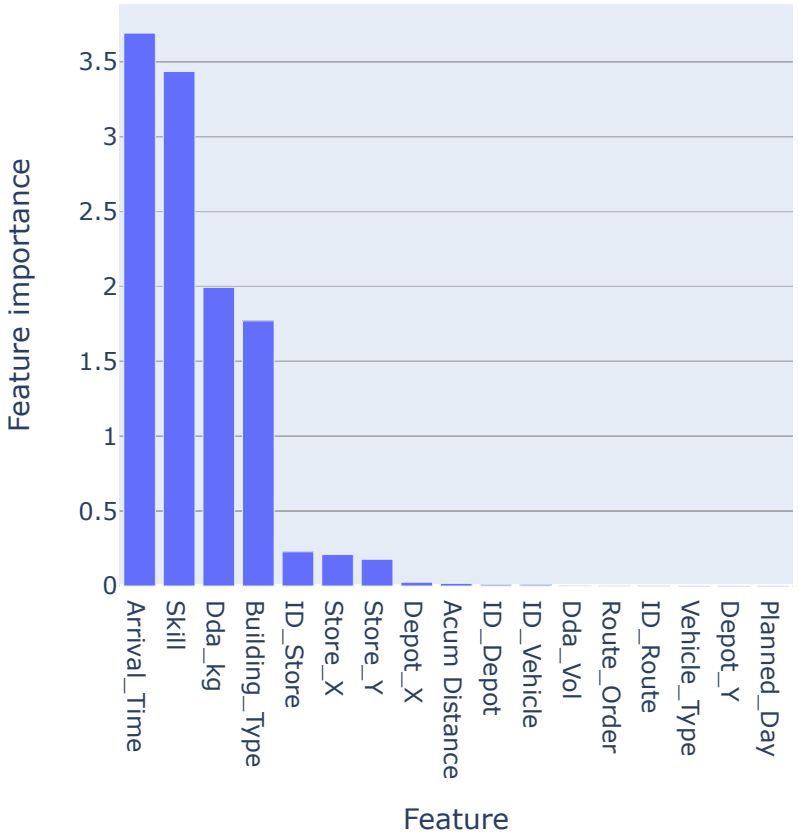
	Algorithm name	Spearman correlation
0	StackEnsemble	0.95306
1	StandardScalerWrapper, LightGBM	0.94014
2	VotingEnsemble	0.93904
3	StandardScalerWrapper, XGBoostRegressor	0.93804
4	MaxAbsScaler, LightGBM	0.92973
5	MaxAbsScaler, LightGBM	0.91384
6	MaxAbsScaler, LightGBM	0.90105
7	StandardScalerWrapper, LightGBM	0.89791
8	MaxAbsScaler, DecisionTree	0.89633
9	StandardScalerWrapper, LightGBM	0.89258
10	StandardScalerWrapper, LightGBM	0.89079
11	MaxAbsScaler, XGBoostRegressor	0.88897
12	StandardScalerWrapper, LightGBM	0.88672
13	MaxAbsScaler, RandomForest	0.88618

Tabla 6. Los mejores algoritmos encontrados por AutoML.

4.2 Modelo

El mejor modelo que encuentra AutoML lo descargamos en formato .pkl para realizar la predicción de los datos de test.

Figura 10. Ranking de las caracteristicas de las entrega con el algoritmo StackEnsemble, Spearman correlation




```
[05:46:10] WARNING: /mnt/xgboost/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squ
arederror.
[05:46:10] WARNING: /mnt/xgboost/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squ
arederror.
```

Para hacer la estimación del tiempo de servicio con el modelo requerimos ingresar los campos que se utilizaron en su construcción.

```
camposRequeridosPrediccion=['ID_Visit', 'Planned_Day', 'ID_Route', 'ID_Store', 'ID_Vehicle',
                            'Route_Order', 'Store_X', 'Store_Y', 'ID_Depot', 'Depot_X', 'Depot_Y',
                            'Dda_kg', 'Dda_Vol', 'Skill', 'Building_Type', 'Acum Distance',
                            'Arrival_Time', 'Vehicle_Type']
```

Realizamos la estimación del tiempo de servicio tanto para los datos Train como para los Test:

```
simpliRoute['Service_Time_Prediction']=model.predict(simpliRoute[[camposRequeridosPrediccion]])

simpliRouteTest['Service_Time_Prediction']=model.predict(simpliRouteTest[[camposRequeridosPrediccion]])
```

Con la estimación en la base de Train hacemos la gráfica de la figura 11 y corroboramos que el modelo hace una buena estimación del tiempo de servicio.

Creamos dos variables para entender las predicciones del modelo:

1. Diferencia del tiempo de servicio real y el predicho, nos sirve para estimar los minutos de diferencia que se obtienen con la estimación del tiempo de servicio. Se encuentra que las estimaciones del tiempo de servicio, tipicamente tienen una discrepancia de ± 1.3 minutos con el valor real.

   ```
   simpliRoute['Service_Time_diff']=simpliRoute['Service_Time']-simpliRoute['Service_Time_Prediction']
   ```

2. diferencia porcentual: es el tiempo de servicio real menos el predicho sobre el tiempo real. No tiene unidades e indica en forma de fracción la desviación del tiempo de servicio estimado con el real. La figura 12 muestra esta cantidad como función del tiempo de servicio. Aunque se obtienen diferencias de hasta el 500% para tiempos de servicio de un minuto, el 80% de los tiempos de servicios son mayores a once minutos, desde donde la discrepancia es menor al 10% entre el valor estimado con el modelo y el real.

   ```
   simpliRoute['Service_Time_diff_percent']=100*np.abs((simpliRoute['Service_Time']-simpliRoute['Service_Time_Prediction'
   ]))/simpliRoute['Service_Time']
   ```

Figura 11. Predicted vs. True

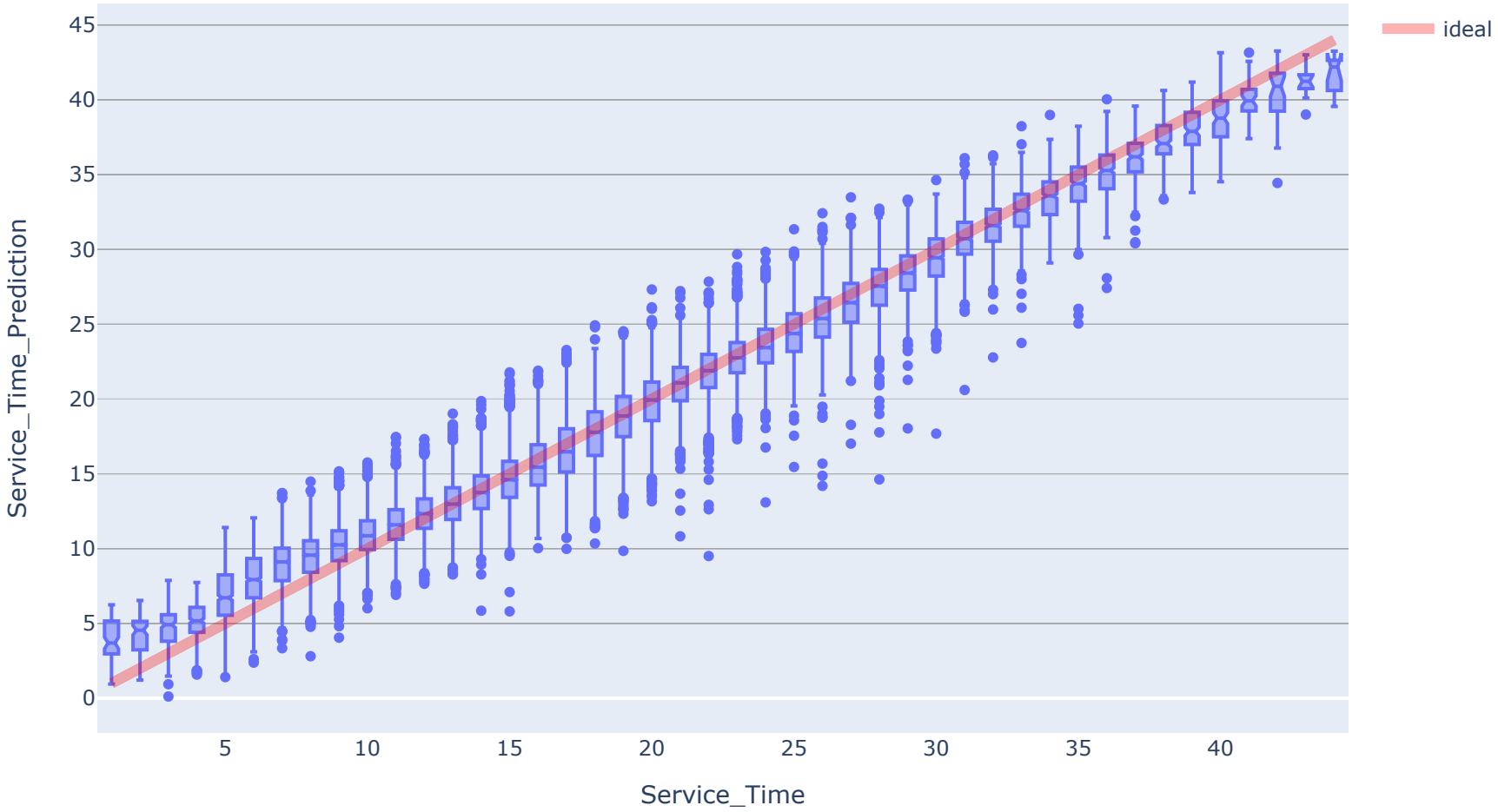


Figura 12. Relative error in the prediction of service tim

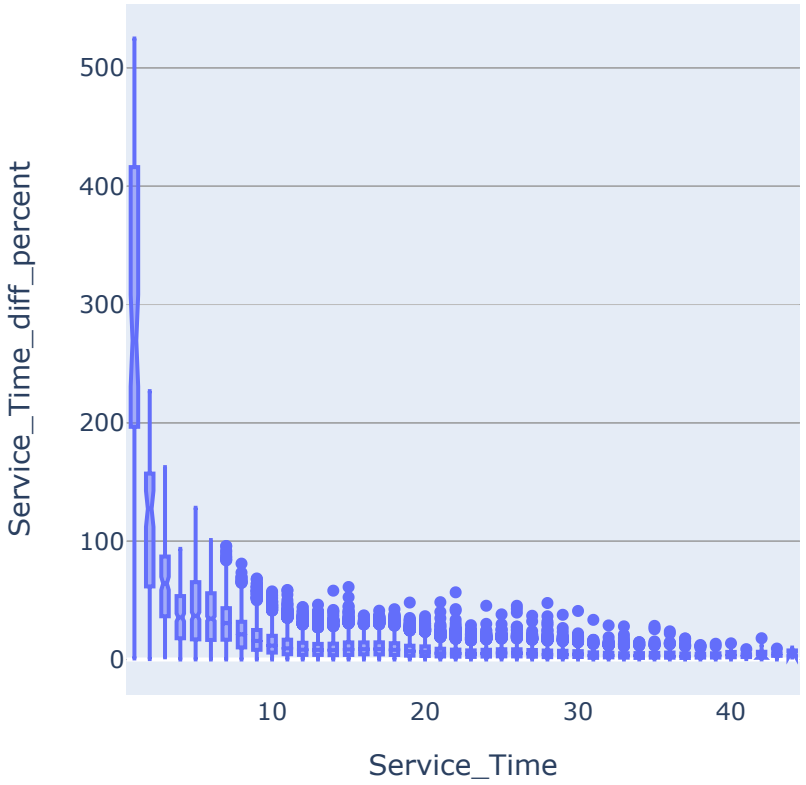
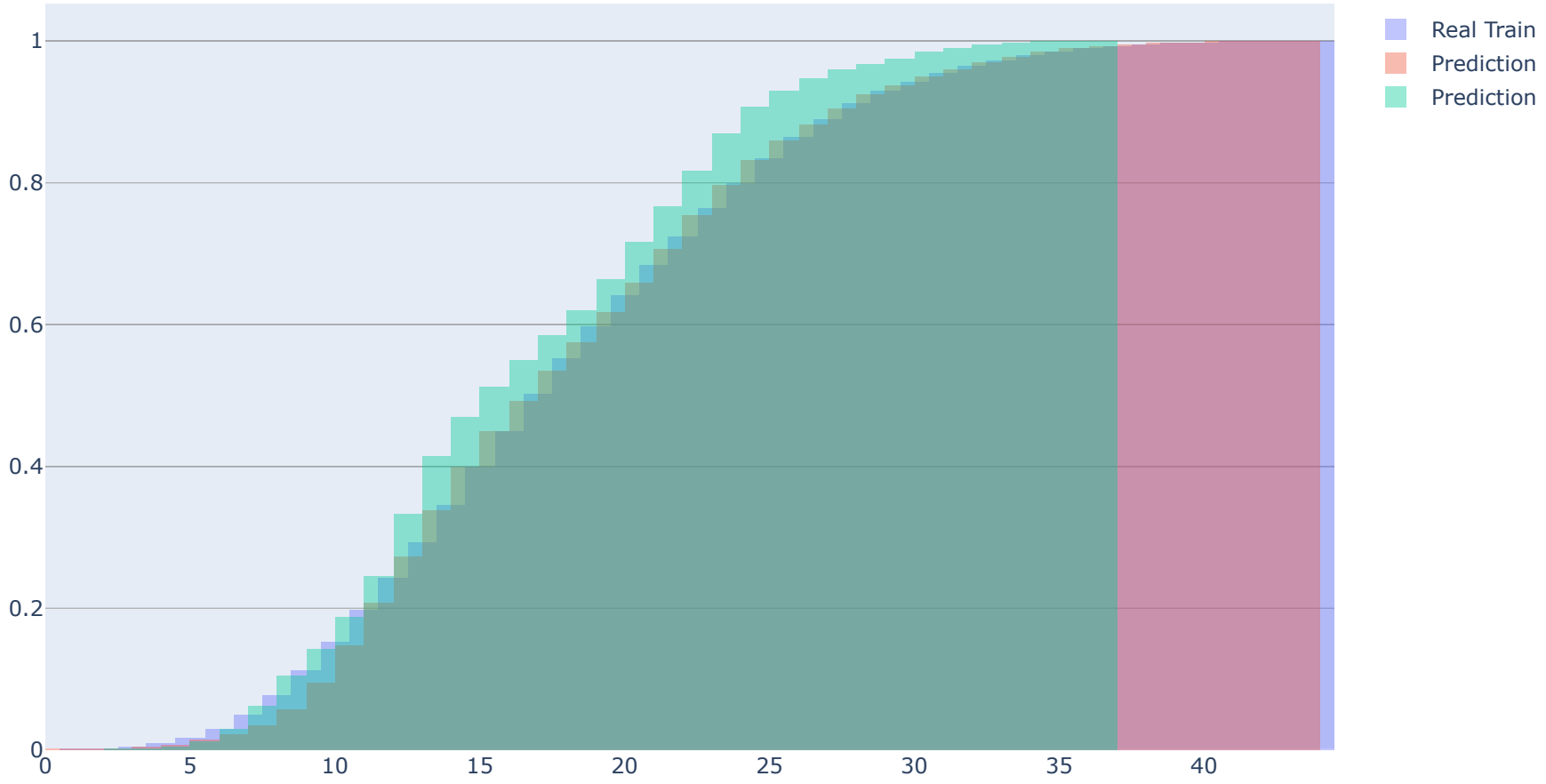


Figura 13. Distribución acumulada de tiempo de servicio



4.3 Visualización de las rutas con el tiempo de servicio estimado en el archivo de test

Utilizamos el mismo método de visualización de rutas para los datos de la base testing usando el tiempo de servicio estimado con el modelo de stack ensemble. Analizando una ruta aleatoria en un día aleatorio vemos el mismo comportamiento que veiamos con los datos del archivo de training: los tiempos de servicio mayores se tienen al entregar los paquetes más pesados y va aumentando a medida que se avanza en la ruta, también se observa que sobre 20 entregas se tiene el mayor tiempo de servicio.

La Visualización de la ruta de un vehiculo en un día particular de la base de testing con el tiempo de servicio estimado con el mejor modelo de Machine Learning entregado por AutoML. Las caracteristicas observadas en las figuras 7 y figuras 8 también se observan en las figuras 14 y figuras 15, dando soporte al buen comportamiento de la estimación del modelo.

Figura 14a. Test data.
Ruta individual de un vehiculo aleatorio:11 un día aleatorio:122

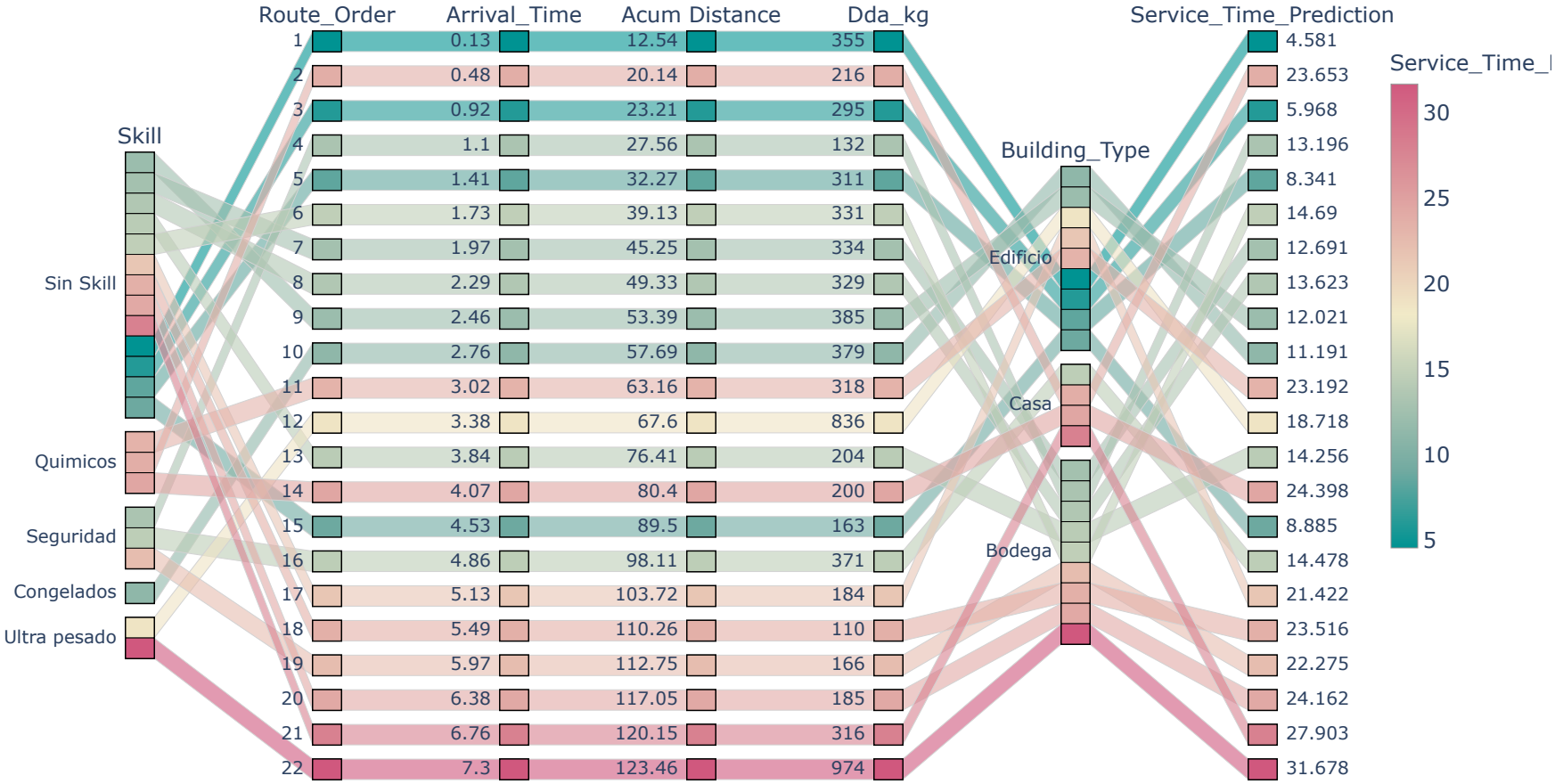


Figura 14b. Test data.
Rutas de diez vehiculos el mismo día de la figura 14a

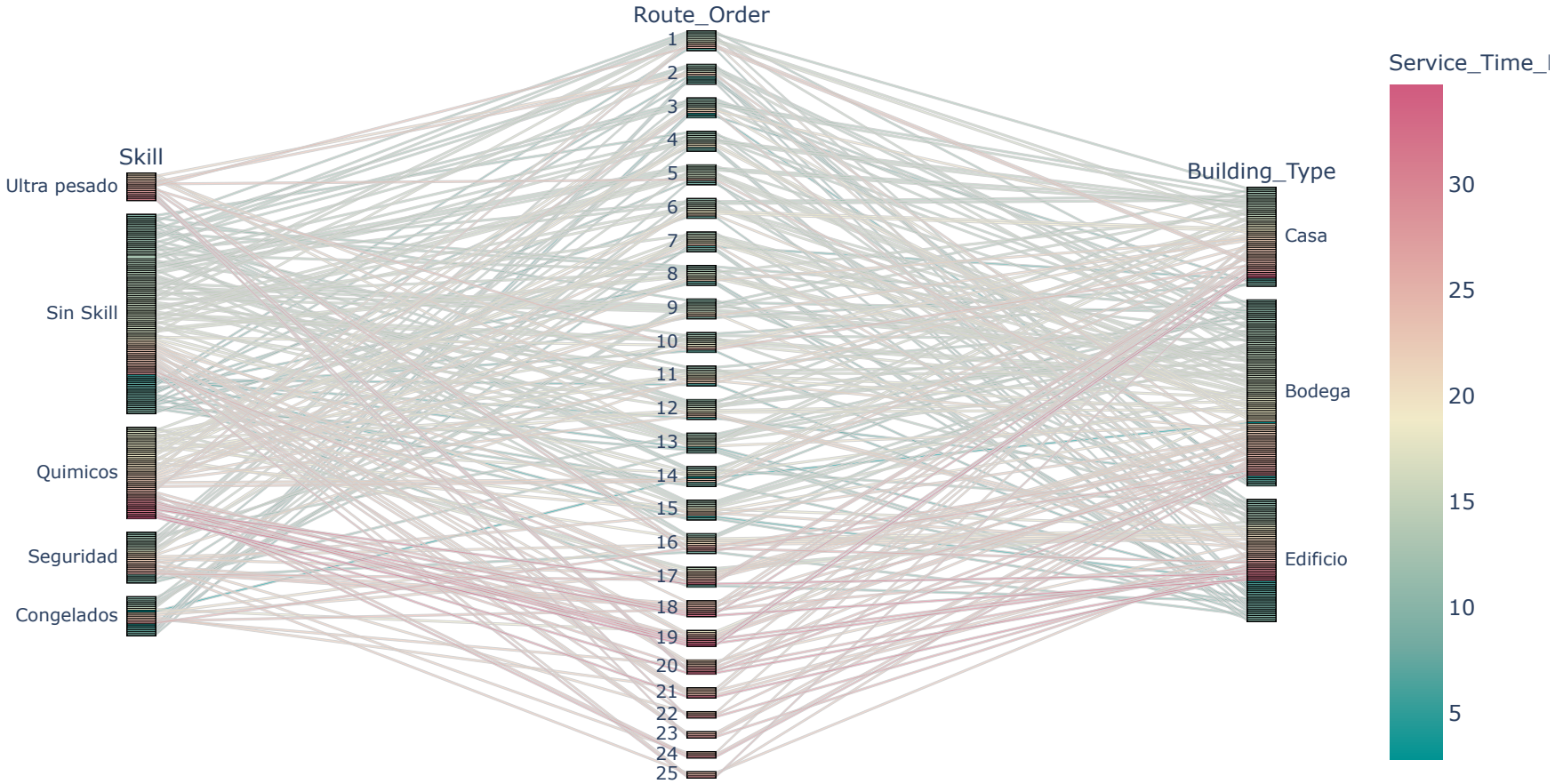


Figura 15a. Testing data. Ruta individual de el vehiculo 11 el día 122

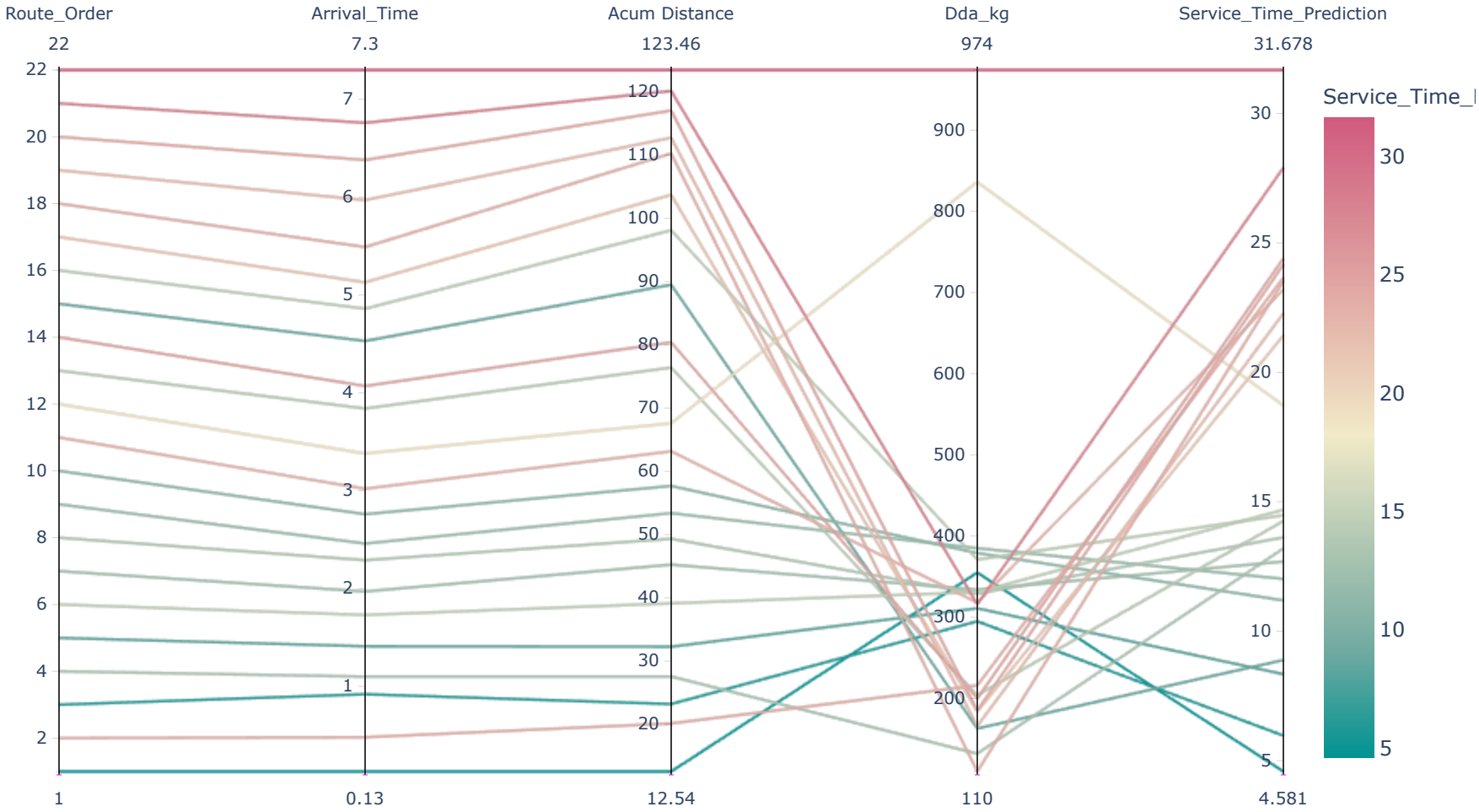
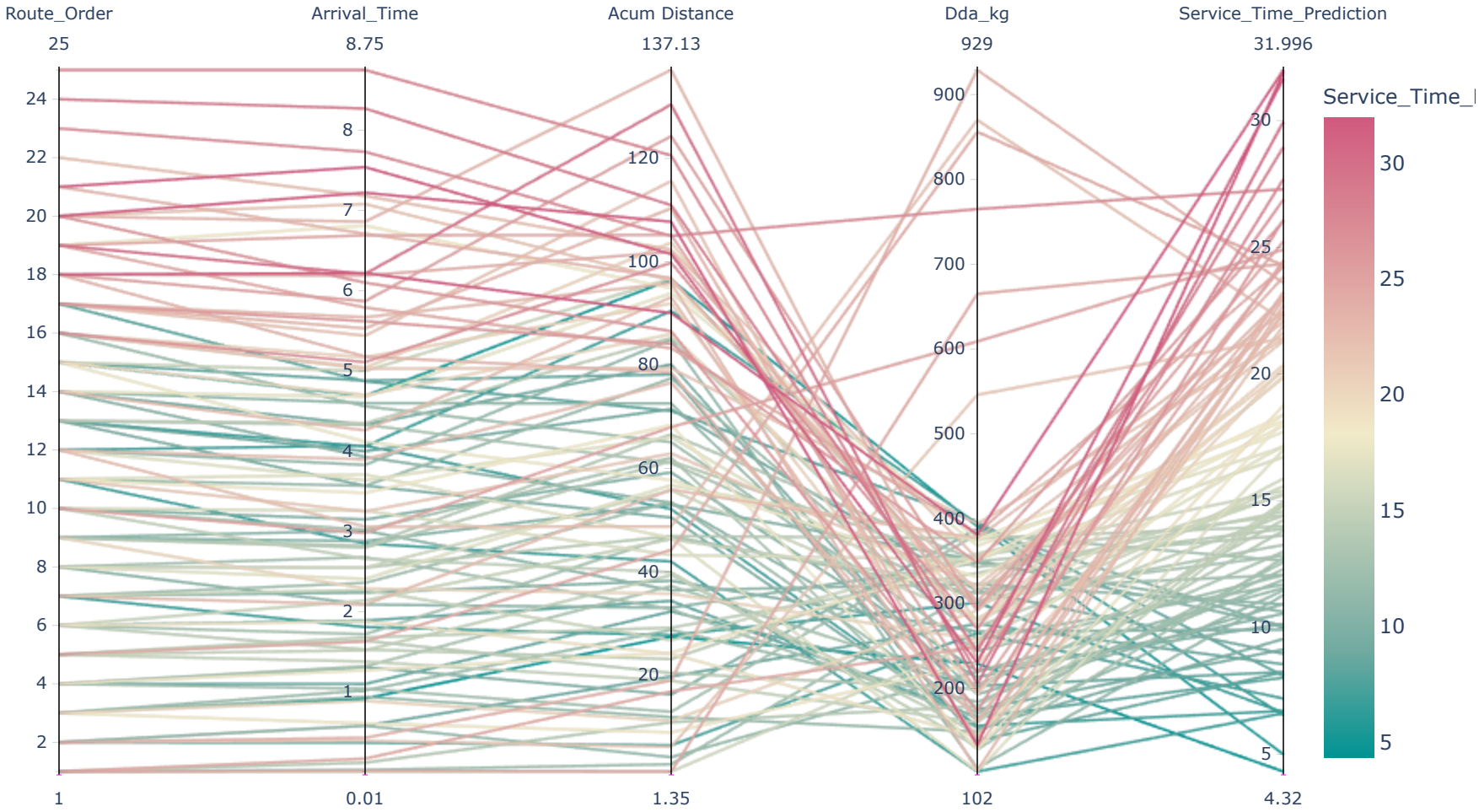


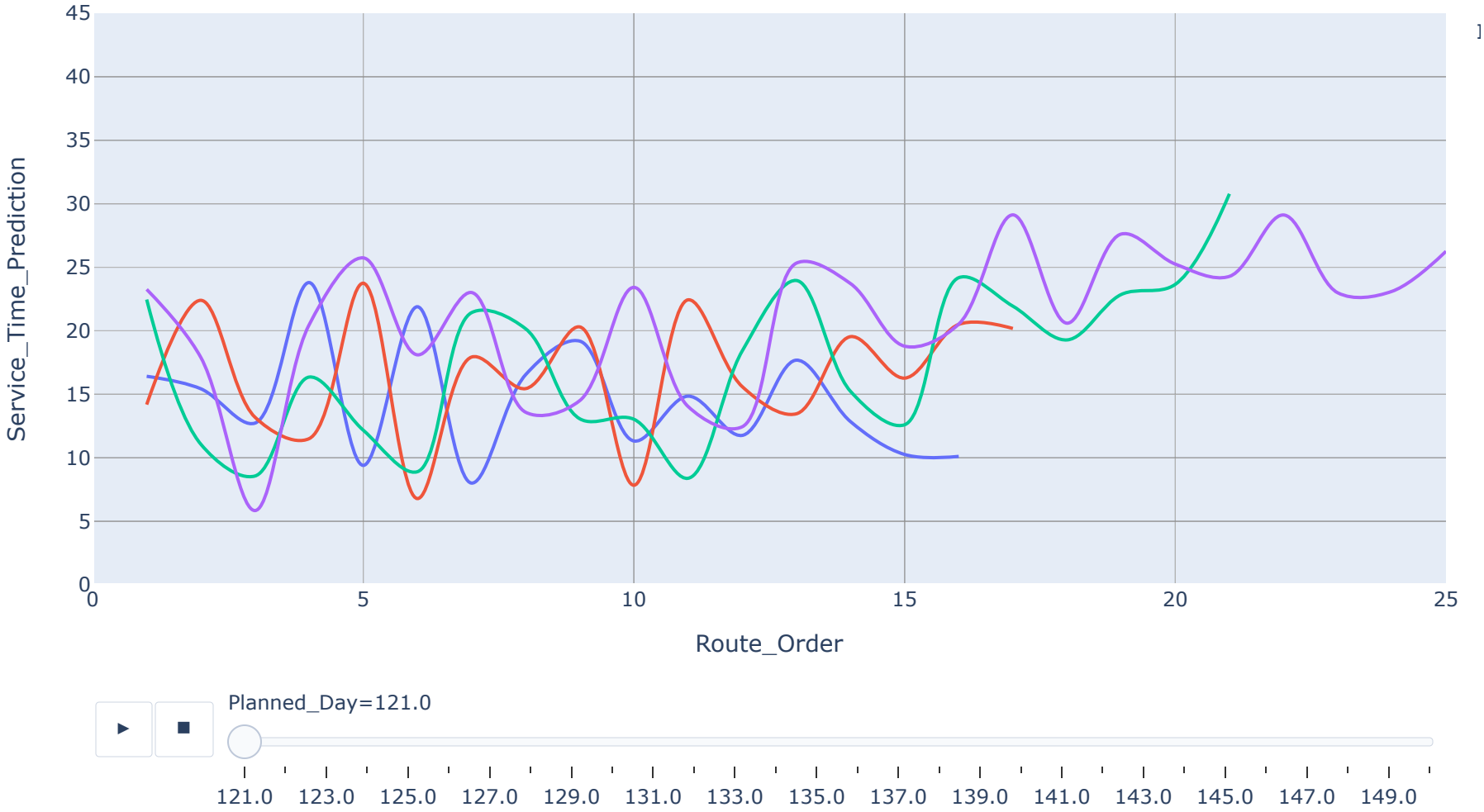
Figura 15b. Testing data. Rutas de diez vehiculos el mismo día de la figura 15a



4.5 Visualización colectiva de rutas

Graficamos las rutas para los mismos vehiculos de la figura 9 todos los días de la base testing usando la estimación del tiempo de servicio hecha con el modelo Stack Ensemble. Se observa un comportamiento similar al de los datos reales en las trazadas de las curvas Link a las curvas reales del vehiculo 4.

Figura 16a. Tiempo de servicio como función del orden de la ruta evolucionando con el día.



Conclusiones

En tiempos de Coronavirus se hace relevante que el contacto físico entre personas se reduzca a su mínima expresión. Los datos de ruteo de vehiculos proporcionados por SimpliRoute (https://www.simpliroute.com/) permiten extraer información relevante para tomar acciones que reduzcan este tiempo en el cual el virus puede ser transmitido entre quien entrega y quien recibe un producto. Con métodos de visualización de datos y machine learning se encuentra que el tiempo de servicio se va incrementando a medida que se avanza en la ruta de entrega, en general las últimas entregas presenta un mayor tiempo de contacto que las primeras. Así mismo, el peso del producto y la habilidad requerida para la entrega juegan un rol capital en la duración de la entrega en cada visita. El servicio de Aprendizaje Automático ó AutoML de Azure Machine Learning (https://azure.microsoft.com/es-es/services/machine-learning/#features) nos provee de una herramienta para poder explorar una gran variedad de algoritmos, junto con sus parámetros, para poder estimar el tiempo de servicio como función de las características de la ruta y el paquete. En base a lo que se queria inicialmente con este análisis podemos concluir:

1. Algunas caracteristicas analizadas permiten estimar el tiempo de servicio en las entregas, principalmente el tiempo de arribo, el skill y el peso del paquete.
2. AutoML encuentra un conjunto de modelos que pueden estimar en buena medida el tiempo de servicio, entre estos el mejor encontrado es un Stack Ensamble, que consiste en considerar weak learners heterogéneos, los que aprenden en paralelo y los combina entrenando un metamodelo para generar una predicción basada en las diferentes predicciones de los modelos débiles. Se obtiene una correlacion Spearman de 0.95, una variación explicada de 0.91 y un score R^2 de 0.91. Alrededor de tiempos de servicios de 15 minutos (los más comunes) la incertidumbre en la predicción es típicamente de 1 y 1.5 minutos. El modelo tiende a predecir más tiempo de servicio para las primeras entregas de la ruta y menos para las del final, pero en general tiene el mismo comportamiento con el resto de variables.
3. Dada la cantidad de datos y lo entremezclados que están es posible perderse y no ver patrones pero con la técnología de inteligencia artificial los podemos resaltar fácilmente. Un correcto análisis de los datos permite descubrir patrones y poder estimar futuros eventos con bajo nivel de incertidumbre, hace que las soluciones que se construyan de cara al cliente tengan mayor valor.
4. Las últimas entregas de la ruta son las que más tiempo de servicio tardan, junto con las de mayor peso. Se recomienda, para reducir el tiempo de servicio a máximo 30 minutos, hacer que las rutas no tengan más de 20 entregas por día y que los paquetes más pesados sean planificados al comienzo de la ruta.

Germán Gómez Vargas
Email:gagomezv@gmail.com
23 de mayo de 2020