Pronostico de costos de garantías a corto plazo para HP Inc.

Reporte Técnico

Daniel Nuño

November, 2022

Indice de contenidos

1 Introducción

Cada mes la organización de finanzas debe proveer un estimado de los gastos mensuales. Actualmente el proceso es intensivo en tiempo y en labor, dejando mucho que desear a la precisión.

1.1 Definición del problema

Mes con mes la organización de finanzas, en conjunto la organización de Customer Support, debe proveer un estimado de los gastos y costos por garantías ejercidas que tendrá en el corto plazo, entre tres y doce meses. La solución actual deja que desear en cuanto la precisión, además de ser muy demandante en tiempo y personal.

Estas garantías son de las computadoras e impresoras de uso comercial y personal vendidos de HP en todo el mundo. Geograficamente comprende 3 regiones y 10 mercados:

- América
  - Norte América
  - Latin América
- Europa, Africa y Medio Oriente
  - Reino unido
  - Europa Central
  - Europa Sur
  - Europa Noreste
  - Africa y Medio Oriente
- Asia Pacifico
  - China
  - Asia Mayor
  - India

En tipo de producto comprende 4 segmentos:

- Computadoras
  - Comercial - Business PC Solutions
  - Consumo - Consumer PC
- Impresoras
  - Comercial - Office Printing Systems
  - Consumo - Home Printing Systems

El objetivo es crear una solución que pueda proveer una precisión, al menos, igual a las soluciones actuales, pero sin la bruma, el trabajo y el tiempo que conlleva hacerlo mes con mes. Idealmente será completamente automática, supervisada, online, pero hay consideraciones que no están capturadas en los datos, como información de partes altamente defectuosas, problemas en la cadena de suministro o inversiones.

Estas estimaciones en conjunto de otra información o estimaciones proporcionado por otras organizaciones tienen tres propositos principales que se usan internamente: - Estimación del

flujo de efectivo. - Estimación de los estados financieros de la empresa. - Responsabilidad a los altos ejecutivos.

Parte de la visión de HP Inc es la innovación digital e internamente transformar la forma en que trabajamos. El métrico principal es la precisión de la predicción evaluado mes con mes, es decir la diferencia entre predicción y real. El benchmark es la precisión de la solución actual. Adicional, métricos relevantes son (1) cuantos días de laborales se puede reducir para la entrega de la predicción. Si ahora tarda un ciclo de 10 días en entregar entonces que tarde menos de 10 días. Y (2) cuantas horas de trabajo se reducen mes con mes, trabajo en horas por trabajador para entregar la predicción de gastos y costos.

Actualmente esta tarea tiene un costo inherte a la labor de todos los que participan que teoricamente puede reducirse con una nueva implementación. La solución no debe canjear precisión por costo, sino que, por lo menos, la precisión debe ser la misma.

Los costos y gastos se reportan mes con mes y se componen de costos regionales, gastos globales, y reservas y amortizaciones. Los costos globales son en su mayoría fijos relacionados a empleados o inversiones. Las reservas y amortizaciones responden a ahorros hechos para cubrir los costos basados en las ventas. Los costos regionales corresponden a costos fijos de empleados, pero también a gastos variables operativos como partes de repuesto, cadena de suministro, logistica, trabajo de ingenieros en la reparación, y llamadas de asistencia.

- Total Warranty Expense
  - Region Owned Expense
    - Variable Expense
      - Contact Center
      - Delivery
      - Supply Chain
      - Other Repair Cost
    - Repair OH Expense
      - Contact Center OH
      - Delivery OH
      - Supply Chain OH
    - Other Warranty Expense
  - Worldwide Owned and Allocated Expense
    - CS HQ Owned and Allocated
      - CS HQ
      - CS Investments
    - GBU Owned and Allocated
      - GBU Owned and Allocated
  - Net Reserve Expense
    - Net Reserve Expense

- Accrual for Shipments
- Amortization

Ver Sección 4.1 para una mayor explicación.

Para gastos de variables de contact center necesitamos saber tres cosas:

- V = Cantidad de unidades vendidas en un periodo.
- L = Porcentaje de productos con fallas o
- V*L = cantidad de asistencias sin reparación.
- C_llamada = costo promedio por llamada.

$$ccvariable = V * L * C_{llamada}$$

Para gastos variables de reparación necesitamos saber tres cosas:

- V = Cantidad de unidades vendidas para un periodo.
- R = Porcentaje de unidades vendidas que necesiten reparación.
- C_ reparación = El costo promedio de reparación.

$$reparacionvariable = V * R * C_{reparacion}$$

Para los costos fijos o over head necesitamos saber dos cosas:

- E = Cantidad de empleados.
- C_empleado = Costo promedio por empleado.

$$overhead = E * C_{empleado}$$

Otro de los requisitos es la granularidad en la geografía (mercado) y el tipo de producto (segmento), lo que agrega complejidad al proceso por que los costos de un segmento y mercado terminan siendo diferentes. Son 10 mercados y 5 segmentos.

Las asunciones hasta ahora son:

- Tiene tendencia.
- Tiene estacionalidad.
- Es autorregresivo.
- Es un proceso estocástico porque hay costos no previstos.
- Los números reportados no son perfectos por errores humanos, cambios operativos, contables y de sistemas.
- Datos más recientes y entendimiento del modelo de negocio son más importantes para los pronosticos al futuro.

- Un modelo explicativo de cada linea de costos es más importante que los datos historicos. El pronóstico a futuro de variables operativas es vital para una buena precisión.
- Backlog es un punto de partida importante para cada mes.

2 Datos

Los datos a utilizar son los costos mensuales de cada una de las variables que componen operativamente la organización de garantías. Estos costos monetarios al ser divididos por mercado y por línea de productos estamos hablando de múltiples series de tiempo. En cuanto al rango, los datos disponibles son de noviembre 2016 a Noviembre 2022.

Los datos financieros son recolectados del General Ledger. Datos operativos son recolectados de diferentes sistemas dependiendo de la región o el tipo de producto. Los datos son consolidados en una base de datos, por lo tanto no se les aplico tratamiento más que etiquetado de datos y codificación de los valores para proteger la confidencialidad de HP.

Estos datos son propiedad y confidenciales de HP Inc. y son usados por mi persona como empleado y bajo guía de mi jefe con la intención de mejorar el proceso.

2.1 Benchmark actual metodología

Primero vamos a analizar el problema y establecer un benchmark, comparando los pronósticos con los resultados reales históricos. Los datos van de la siguiente manera:

Para cada mes existen n estimaciones de costos pasados, que pueden ser expresados como un vector:

$$\begin{aligned} C &= \text{costo} \\ t &= \text{periodo} \\ Ca_t &= \text{costo subindice t, costo del periodo} \\ Cf_t &= \text{costo subindice t, costo del periodo} \\ n &= \text{número de periodos pasados} \end{aligned}$$

$$flash = \{Cf_{t-1}, Cf_{t-2}, Cf_{t-3}, \dots, Cf_{t-n}\}$$

Y el valor real, también llamado *actual*

$$actual = Ca_t$$

El vector de error o desviación para cada periodo sea la diferencia del valor actual y cada uno de los valores del vector flash sobre el valor actual.

$$error = \{\frac{Ca_t}{Cf_{t-1}} - 1, \frac{Ca_t}{Cf_{t-2}} - 1, \dots, \frac{Ca_t}{Cf_{t-n}} - 1\}$$

De aquí podemos calcular el valor esperado y desviación estándar del error, lo cual determina nuestro benchmark.

De forma matricial, cada fila es un periodo de la forma que incluye el costo real y cada uno de las estimaciones pasadas:

$$\{actual, flash\}$$

$$\{Ca_t, Cf_{t-1}, Cf_{t-2}, Cf_{t-3}, \dots, Cf_{t-n}\}$$

Definiendo $n = 6$ obtenemos la siguiente matriz.

	market	line_cost	month	t	t-6	t-5	
0	Latin America Market	Supply Chain	2022-03-01	3722255.65	4.183472e+06	4.183472e+06	3.7638
1	Latin America Market	Supply Chain	2022-04-01	2678478.32	4.189132e+06	3.424980e+06	3.3657
2	Latin America Market	Supply Chain	2022-05-01	3278974.45	3.680119e+06	3.680119e+06	3.6801
3	Latin America Market	Supply Chain	2022-06-01	2979226.67	3.701588e+06	3.701588e+06	3.6993
4	Latin America Market	Supply Chain	2022-07-01	2739678.74	3.708447e+06	3.694892e+06	3.5146

market	line_cost	t-1	t-2	t-3	t-4	
Latin America Market	CS HQ Owned and Allocated	0.003791	0.053210	0.001957	0.011855	0.01
	Contact Center Expense	-0.124099	-0.150918	-0.168062	-0.201347	-0.24
	Contact Center OH	0.196107	0.245507	0.219371	0.133714	0.66
	Delivery	-0.050135	-0.058109	-0.066143	-0.076324	-0.07
	Delivery OH	0.052647	0.034742	0.112444	-0.010762	-0.02
	GBU Owned and Allocated	-0.055085	-0.098478	-0.059607	-0.027108	-0.02
	Supply Chain	-0.079935	-0.080191	-0.078213	-0.092346	-0.13
	Supply Chain OH	-0.127181	-0.055031	-0.170837	-0.300577	-0.38
North America Market	CS HQ Owned and Allocated	0.005928	0.086463	0.019931	0.028349	0.03
	Contact Center Expense	-0.009535	-0.015364	-0.020528	-0.071673	-0.11
	Contact Center OH	0.084855	0.101132	0.086595	0.031700	0.02
	Delivery	-0.008201	-0.001647	-0.020985	-0.043130	-0.08
	Delivery OH	-0.070620	-0.072191	-0.073141	-0.101115	-0.15
	GBU Owned and Allocated	-0.133039	-0.281934	-0.212762	-0.156756	-0.14
	Supply Chain	-0.032593	-0.033972	-0.057142	-0.098773	-0.16
	Supply Chain OH	-0.067556	-0.201321	-0.219529	-0.211630	-0.26

market	line_cost	t-1	t-2	t-3	t-4	t-
Latin America Market	CS HQ Owned and Allocated	0.018333	0.147832	0.143888	0.144439	0.15459
	Contact Center Expense	0.130854	0.109480	0.103554	0.105547	0.09698
	Contact Center OH	0.436545	0.422398	0.404777	0.463803	1.39229
	Delivery	0.046175	0.050776	0.062197	0.062856	0.05624
	Delivery OH	0.061796	0.094700	0.115666	0.235778	0.44574
	GBU Owned and Allocated	0.103451	0.093697	0.100340	0.146038	0.14161
	Supply Chain	0.128471	0.139951	0.125363	0.151937	0.11569
	Supply Chain OH	0.363672	0.414622	0.385338	0.419413	0.43043
North America Market	CS HQ Owned and Allocated	0.019448	0.200881	0.140753	0.139666	0.15087
	Contact Center Expense	0.059857	0.063589	0.078231	0.062779	0.05437
	Contact Center OH	0.430190	0.427675	0.397990	0.392990	0.35741
	Delivery	0.076977	0.087451	0.085637	0.098303	0.10277
	Delivery OH	0.092196	0.090903	0.118212	0.088231	0.16208
	GBU Owned and Allocated	0.140639	0.082510	0.070829	0.127551	0.12819
	Supply Chain	0.142980	0.159445	0.154066	0.170781	0.13556
	Supply Chain OH	0.200087	0.356826	0.345518	0.372685	0.33384

Como ya se esperaba, mientras más periodos hacía el futuro y más alejado de *t*, mayor es el error esperado. Mientras más cerca mejor es el pronóstico.

- La media en *t-6* es entre -0.46 a 0.02. Desviación estándar entre 0.1 a 0.95.
- En *t-3* entre -0.21 a 0.001, y desviación estándar entre 0.07 a 0.4.
- En *t-1* entre -0.009 a 0.19, y desviación estándar entre 0.01 a 0.4.

Sorpresivamente CS HQ Owned and Allocated es bastante buenos en el pronóstico. En la mayoría de los casos, lo estimado esta subestimado y los costos reales son mayores.

Como ya se había comentados anteriormente, el problema se magnifica cuando más específico en el detalle y concentras en una sub línea de costo o en un tipo de producto.

Despues de varias semanas (20 hrs.) discutiendo con expertos de finanzas, de operaciones y explorando los datos, se ha llegado a dos conclusiones que ya se asumian posibles.

El transcurso de 5 años se han pasado varios cambios operacionales que afectan la estructura de costos, la estrategia de atención y reparación para la region o tipo de producto, uso de proveedores tercearios que afecta los costos de diferente manera. La estructura de costos, los datos, siguen cambios operacionales y propias de la region o tipo de producto, lo cual hace el análisis histórico y complicado ya que existe poco información o comentarios que explique estos cambios.

Cambios en los sistemas financieros, la separación de HP, cambios en las lineas de producto y agrupación de regiones, rotación de analistas financieros y cambios en los procesos contables, y errores de los analistas financieros (por que muchos de los costos son registrados por estos)

existen diferencias y atípicos en los datos a traves de los años, y dependiendo de la region y tipo de producto también.

Los datos, los costos, la estructura contable termina siendo imperfecta y poco útil para modelos que dependan 100% de los datos, en este caso datos históricos.

- Los gastos Delivery OH y Supply Chain OH son nuevos. Empezaron a registrarse en noviembre de 2020 entonces solo hay 24 observaciones.
- Delivery sub tipos de costos, llamados Direct e Indirect, empezaron a registrarse en noviembre de 2020 entonces solo hay 24 observaciones.
- Tipo de productos de consumer no registra Delivery por que por el tipo de repación no es necesaria.
- CS HQ Owned and Allocated sub tipo de gastos, llamados IT POA y Rapid and Radical, solo ocurren una vez por mes.
- Los costos de Contact Center por mucho años estuvieron rezagados por un mes. En t se registraba la actividad correspondiente pero en $t+1$ los costos. No estaban en par pero desde hace seis meses ahora ya estan alineados desde hace 6 meses.
- Los cartuchos de tinta y toner son un tipo de producto llamados Supplies y existen para impresoras comerciales y consumo (uso personal). En nuestra organización no existe supplies comercial pero sí para consumo. Todos los costos y metricas relacionados a supplies comercial no tienen sentido y tienen que ser eliminados para reducir el ruido.
- Otros negocios tiene su propio soporte de garantias.
- Porque somos una compañia y existen economias de escala, en su mayoria todos los costos son centralizados, independientemente del tipo de producto o país.

2.2 Análisis exploratorio

Por estas razones, la estrategia para resolver el problema sera definir que costos y aque nivel son necesarios a pronosticar, unicamente para dos regiones y sin distinguir por el tipo de producto.

- Latin América / North América
  - CS HQ Owned and Allocated
  - Contact Center Expense
  - Contact Center OH
  - Delivery
  - Delivery OH
  - GBU Owned and Allocated
  - Supply Chain
  - Supply Chain OH

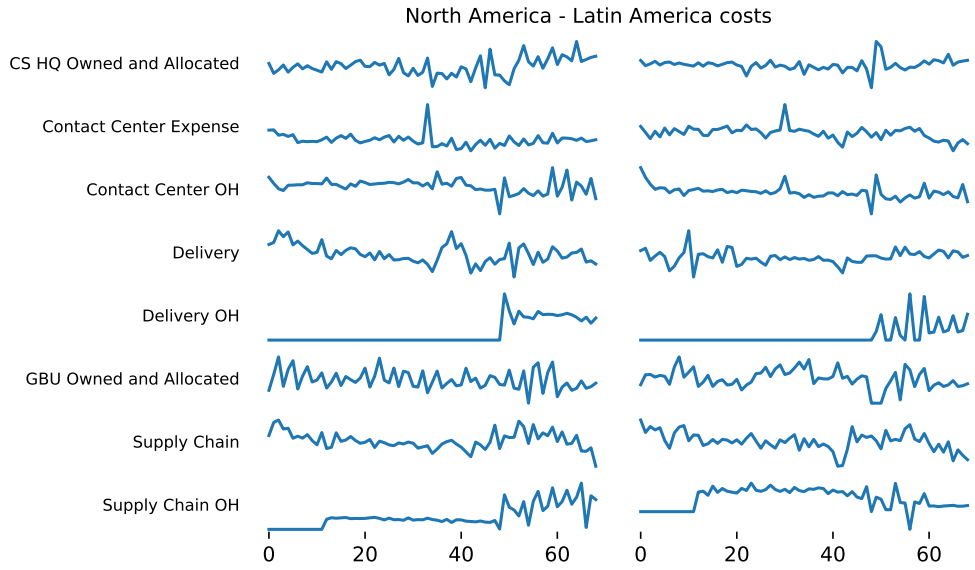
Los datos disponibles son 72 periodos, desde noviembre 2016 hasta octubre 2022. El periodo fiscal para HP comienza en noviembre. Todos los valores son númericos monetarios que representan el gasto o costos incurridos en el periodo.

Al final, para tener congruencia con los datos en cuanto a sub costos, paises, y tipo de producto, cada pronostico realizado para un periodo hacia delante tiene que generar 640 puntos de datos. Por ahora, la estrategía para calcularlo es tomar el valor obtenido y multiplicarlo por una matriz de porcentajes.

Para este análisis me interesa observar lo siguiente para cada serie de tiempo (cada linea de costo sea una serie de tiempo):

- gráficas de serie de tiempo.
- graficas distribucion de valores.
- valores atípicos y nulos.
- varianza.
- sesgo.
- conteo de observaciones y rango de periodos.
- correlaciones.
- tendencia y estacionaridad.
- estacionalidad y ciclos.
- homocedasticidad.
- normalidad.

El análisis sera por línea de costo (8) y mercado (2) reservando los últimos tres meses para hacer pruebas



Se puede apreciar los valores nulos y los atípicos.

- Delivery OH empieza hasta después del mes 48.
- Supply Chain OH empieza hasta después del mes 48.
- Casi todos los valores atípicos son precedidos de un valor nulo (0), significa que en el valor nulo no existe registro de gastos y por lo tanto el siguiente mes es mucho más alto. Esencialmente lo correspondiente del mes pasado más el mes actual. Un mes no hay gasto registrado y al siguiente es lo correspondiente a dos meses. Son dos valores atípicos para tratar por error humano. Como en la serie de CS HQ Owned en Latin América en el mes 50.
- La gráfica muestra que las series de tiempo carecen de gran tendencia, y son cíclicas. Procesos estacionarios y estacionales.

Para la serie de tiempo se usarán variables dummies para considerar los valores atípicos o reemplazarlos por el valor inmediato anterior.

line_cost	NA coeficiente variación	LA coeficiente varianción
CS HQ Owned and Allocated	0.181132	0.218788
Contact Center Expense	0.180603	0.139662
Contact Center OH	0.221789	0.265441
Delivery	0.120434	0.225198
Delivery OH	0.240928	0.700224
GBU Owned and Allocated	0.285684	0.319551
Supply Chain	0.132234	0.133093
Supply Chain OH	0.684712	0.532170

El coeficiente de variación indica que los costos Delivery OH y Supply Chain OH son los más dispersos. Pueden ser los más complicados de pronosticar.

line_cost	CS HQ Owned and Allocated	Contact Center Expense	Contact Center OH	Delivery
count	6.900000e+01	6.900000e+01	68.000000	6.900000e+01
mean	1.897886e+06	5.342644e+06	531078.058376	2.464819e+06
std	3.462872e+05	9.719670e+05	118662.930224	2.990222e+05
min	1.081765e+06	3.614842e+06	259757.510000	1.737022e+06
25%	1.675260e+06	4.870979e+06	466403.175175	2.275964e+06
50%	1.889849e+06	5.288916e+06	549634.680000	2.447753e+06
75%	2.084538e+06	5.711187e+06	583998.585000	2.657105e+06
max	2.854893e+06	1.100208e+07	868604.040000	3.197913e+06

line_cost	CS HQ Owned and Allocated	Contact Center Expense	Contact Center OH	Delivery
count	68.000000	6.900000e+01	68.000000	69.000000
mean	464007.300880	8.357180e+05	111613.533465	415451.227893
std	102274.236981	1.175729e+05	29847.080644	94244.084012
min	235270.406994	5.066162e+05	59912.950000	134312.865192
25%	408129.760845	7.708853e+05	93190.881099	373458.620431
50%	454001.027526	8.609047e+05	108287.055000	409929.900948
75%	503366.920601	9.025187e+05	120293.043207	459638.860000
max	925584.770000	1.364132e+06	233716.999351	789647.231133

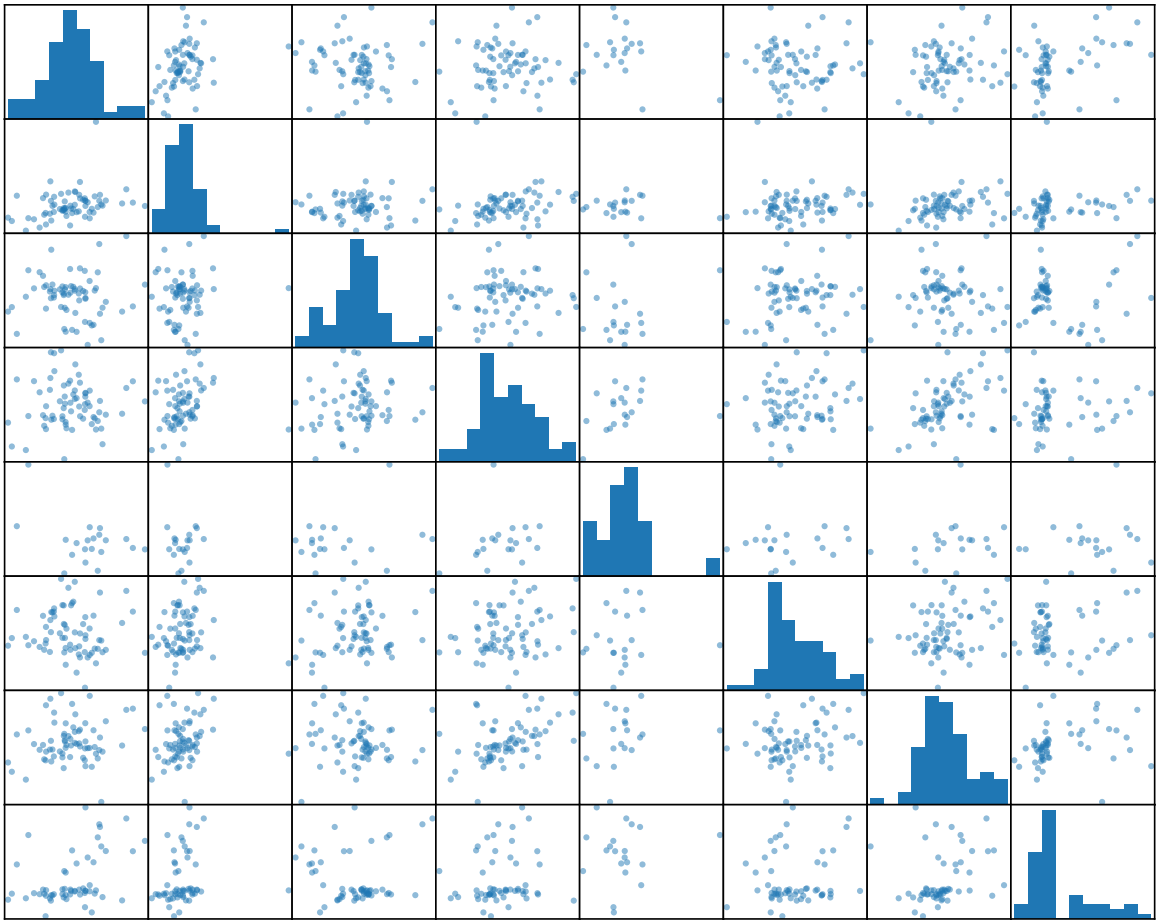
Lo remarcable en la descripción de los datos es:

- La falta de un dato en la serie de costos Contact Center OH.
- La falta de 13 meses en la serie de costos Supply Chain OH. Por las gráficas son datos de finales del 2016.
- GBU Owned and CS HQ Owned para Latin América le faltan 3 meses.
- Las serie de Delivery OH empieza apenas hace 20 meses para Norte América y 15 meses para Latin América.
- Los órdenes de magnitud son muy diferentes entre Latin América y Norte América y entre cada tipo de costo. La covarianza no nos diría mucho y la mayoría de las veces será positiva.
- El coeficiente de variación indica que los costos Delivery OH y Supply Chain OH son los más variables.
- Las series de Delivery OH y Supply Chain OH se tratarán como que comenzaron en noviembre de 2020.

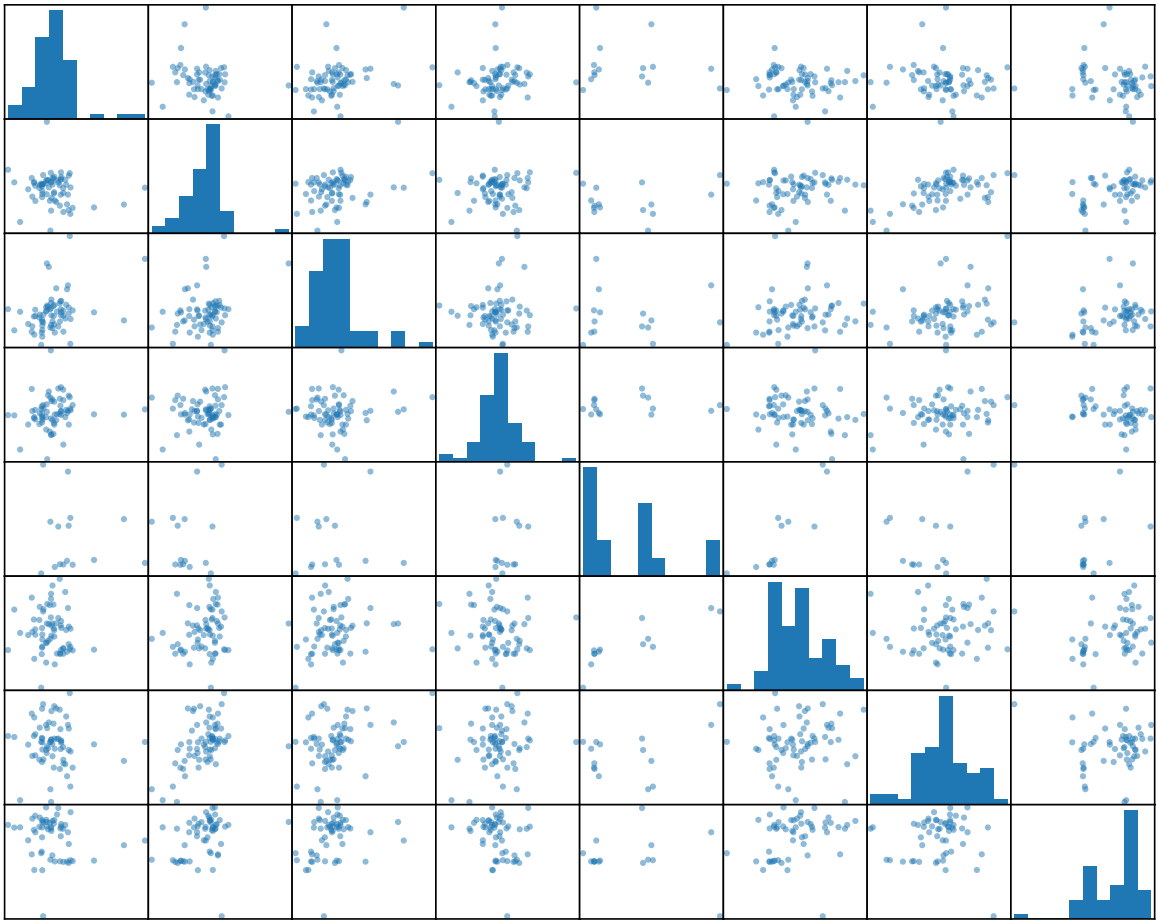
Las distribuciones muestran los atípicos y en su mayoría son sesgados positivamente. Solo dos sesgos negativos en Latin América, 6 sesgos mayores a 1 entre ambas regiones. Las correlaciones entre los tipos de gastos son bajas. En general, para la región de Norte América los costos parecen más correlacionados que para Latin América, donde casi todos son entre -0.2 y 0.2.

En algunos casos como Supply Chain OH y CS HQ Norte América tiene una correlación de 0.5. Supply Chain y Delivery, que esperas que tenga sentido porque ambos son costos variables derivados del volumen de reparaciones hechas. Para Latin América, GBU Owned y Delivery OH tienen una correlación de alrededor de 0.8.

Al final, estas correlaciones y las variables como predictoras de otras variables son inusables para predecir al futuro porque no existen en el futuro entonces no es posible usarse unas a otras. Para este problema usaremos otras series de tiempo pronosticadas, que incluyen métricas operativas, envío de productos terminados e ingresos monetarios.



line_cost	NA sesgo	LA sesgo
CS HQ Owned and Allocated	0.139127	1.764433
Contact Center Expense	2.874702	0.672939
Contact Center OH	0.120663	1.685349
Delivery	0.259752	0.360326
Delivery OH	2.080864	1.039922
GBU Owned and Allocated	0.364877	0.232233
Supply Chain	0.233576	-0.345080
Supply Chain OH	1.282627	-1.315734



2.3 Procesamiento de datos

2.3.1 Atípicos

Ya que los datos que tenemos son series de tiempo no correlacionadas ni dependientes (hasta ahora todos son independientes), los datos son vectores unidemensionales. Y por eso, atípicos sea **3 veces el rango interquartilico** más menos la media. Consideramos que los valor atípicos son usualmente errores humanos y no algo que tiene que ser estudiado a deteminiemto, o que la data que posiblemente explique el fenomeno no esta en este conjunto de datos. Valores nulos después de haber comenzado la serie son también valores atípicos.

```
CS HQ Owned and Allocated tiene 0 valores extremos.
Contact Center Expense tiene 1 valores extremos.
Contact Center OH tiene 0 valores extremos.
```

```
Delivery tiene 0 valores extremos.
GBU Owned and Allocated tiene 0 valores extremos.
Supply Chain tiene 0 valores extremos.
CS HQ Owned and Allocated tiene 2 valores extremos.
Contact Center Expense tiene 1 valores extremos.
Contact Center OH tiene 1 valores extremos.
Delivery tiene 1 valores extremos.
GBU Owned and Allocated tiene 0 valores extremos.
Supply Chain tiene 0 valores extremos.


Delivery OH tiene 1 valores extremos.
Supply Chain OH tiene 0 valores extremos.
Delivery OH tiene 0 valores extremos.
Supply Chain OH tiene 2 valores extremos.
```

2.3.2 Tratamiento de sesgo y estabilización de varianza.

Usando la generalización box-cox ya que los valores son estrictamente positivos. La clase PowerTransformer de sklearn encuentra el mejor lambda y normaliza media cero y desviación estandar unitaria.

```
[0.82336719 0.86976929 0.8549776  0.35878591 2.30165557 0.60703014
 0.64555082 0.97333887]
```

line_cost	0
CS HQ Owned and Allocated	0.019165
Contact Center Expense	0.000781
Contact Center OH	0.022436
Delivery	0.003213
Delivery OH	-0.142155
GBU Owned and Allocated	0.015538
Supply Chain	0.029473
Supply Chain OH	-0.577694

```
[ 1.42885098  3.96265743 -0.18349226  1.73744394 -0.44252247  0.86033292
  1.73344112  0.41355696]
```

line_cost	0
CS HQ Owned and Allocated	0.045933
Contact Center Expense	-0.147578
Contact Center OH	-0.013009
Delivery	0.011757
Delivery OH	0.125001
GBU Owned and Allocated	0.056920
Supply Chain	0.018750
Supply Chain OH	-0.048672

2.3.3 Descomposición series de tiempo (STL)

Hasta aquí los datos que serán usados para entrenar están sin atípicos, transformados y normalizados. Lo siguiente, para modelos de series de tiempo, es estudiar la estacionalidad y estacionariedad con la descomposición de series de tiempo.

Creo que el mejor tipo de modelo sería aditivo por que los valores no cambian mucho con el tiempo, un modelo aditivo es lineal donde los cambios a lo largo del tiempo se realizan consistentemente en la misma cantidad. Una tendencia lineal es una línea recta. Una estacionalidad lineal tiene la misma frecuencia (ancho de ciclos) y amplitud (alto de ciclos).

A diferencia el tipo multiplicativo es no lineal, como cuadrático o exponencial. Los cambios aumentan o disminuyen con el tiempo. Una tendencia no lineal es una línea curva. Una estacionalidad no lineal tiene una frecuencia y/o amplitud creciente o decreciente a lo largo del tiempo.

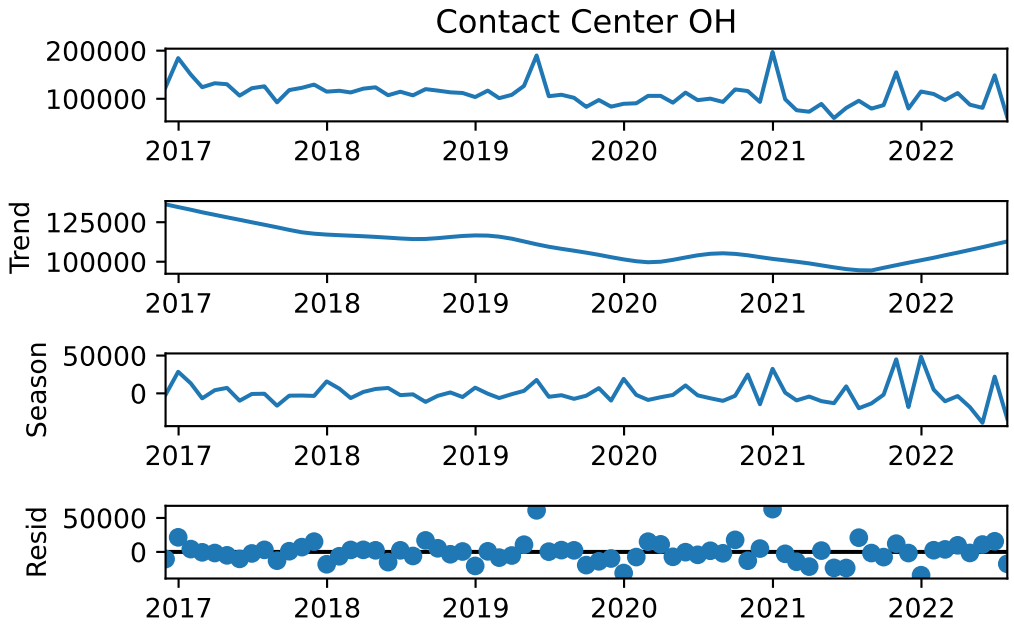
```
CS HQ Owned and Allocated en promedio el residuo representa -0.0164
Contact Center Expense en promedio el residuo representa -0.0071
Contact Center OH en promedio el residuo representa -0.0196
Delivery en promedio el residuo representa -0.0074
GBU Owned and Allocated en promedio el residuo representa -0.0288
Supply Chain en promedio el residuo representa -0.0036
```

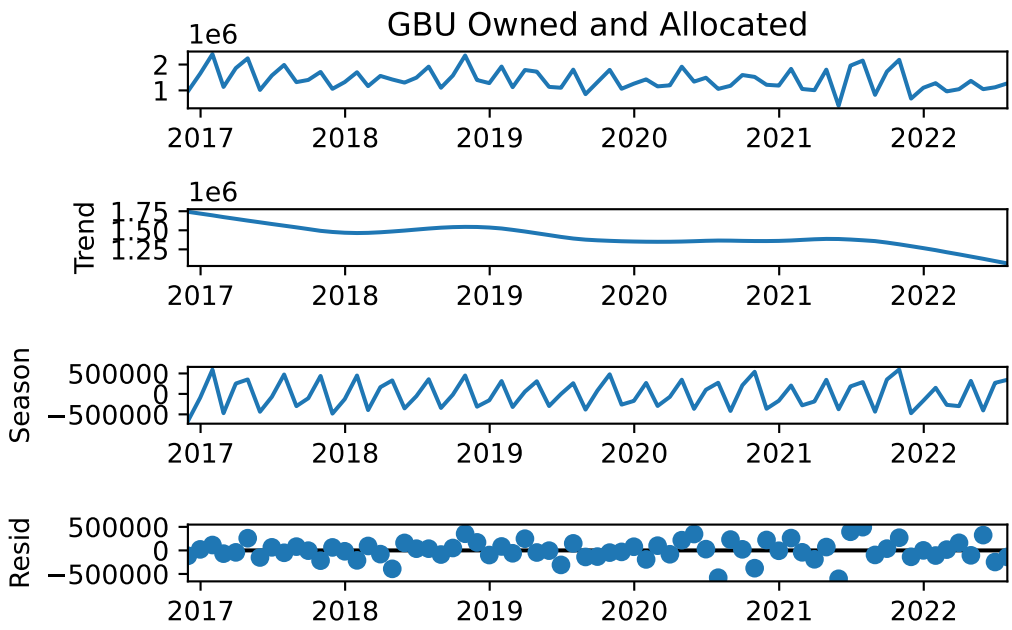
```
CS HQ Owned and Allocated en promedio el residuo representa -0.017
Contact Center Expense en promedio el residuo representa -0.0007
Contact Center OH en promedio el residuo representa -0.0245
Delivery en promedio el residuo representa -0.0279
GBU Owned and Allocated en promedio el residuo representa -0.0776
Supply Chain en promedio el residuo representa -0.0057
```

Debido que a que son muchas gráficas y series de tiempo, quiero calcular la influencia media de los residuos sobre los valores observados con el propósito de comparar entre las series y observar cuales tienen mayores valores "inexplicables" en promedio.

Como podemos observar en los resultados anteriores, *Contact Center OH* tiene un residuo muy alto promedio. *Contact Center Expense* en Latino América parece ser explicado de buena manera por la tendencia y estacionalidad.

GBU Owned and Allocated en ambas regiones tiene una periodicidad muy clara pero también tiene residuos, en promedio, muy altos. Ejemplos gráficos:





3 Modelos

Para transformar la serie de tiempo a un problema de regresión, se planteó un modelo dinámico usando tres rezagos de la misma serie. Además, se introducen otras series de tiempo que tendrían asociación con los costos:

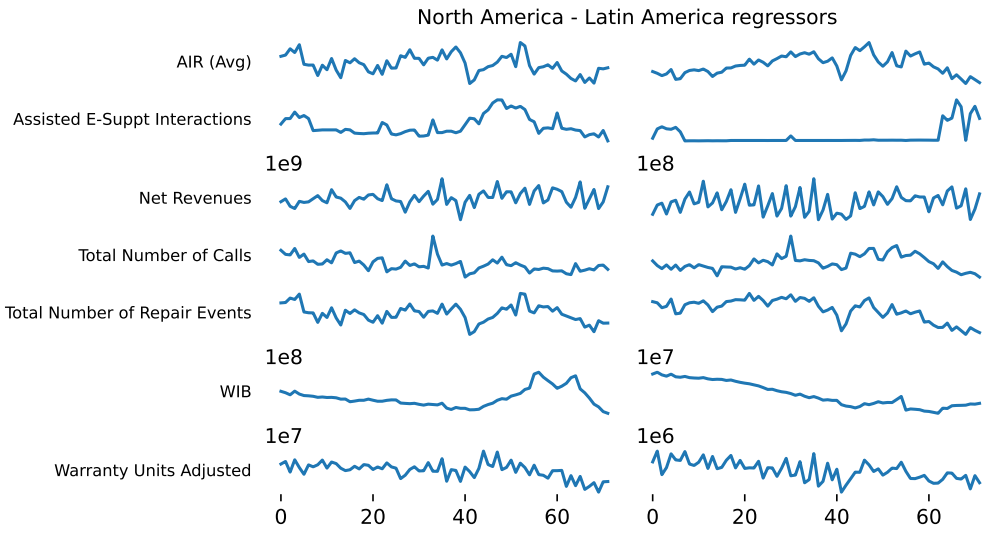
- Ingresos.
- Número de llamadas.
- Número de reparaciones.
- Unidades vendidas.
- Base instalada de unidades en garantía.
- Porcentaje de reparaciones anualizado.

Dichos datos pasan por la misma limpieza:

- Limpieza de atípicos
- Transformación box-cox

Estas series de tiempo tiene la ventaja que son pronosticadas antes de tener que pronosticar los costos que se están estudiando. Es decir, los valores futuros ya son conocidos y por eso sea decidió usarlas. La intención de la predicción es de varios meses hacia delante, en este la propuesta de 3 meses. Es por eso que para t+2 se usa lo estimado por el modelo en t+1 de forma recursiva. Para t+3 se usaría lo estimado por el modelo en t+2.

Las pruebas se realizaron con los modelos de redes neuronales y random forest para regresión.

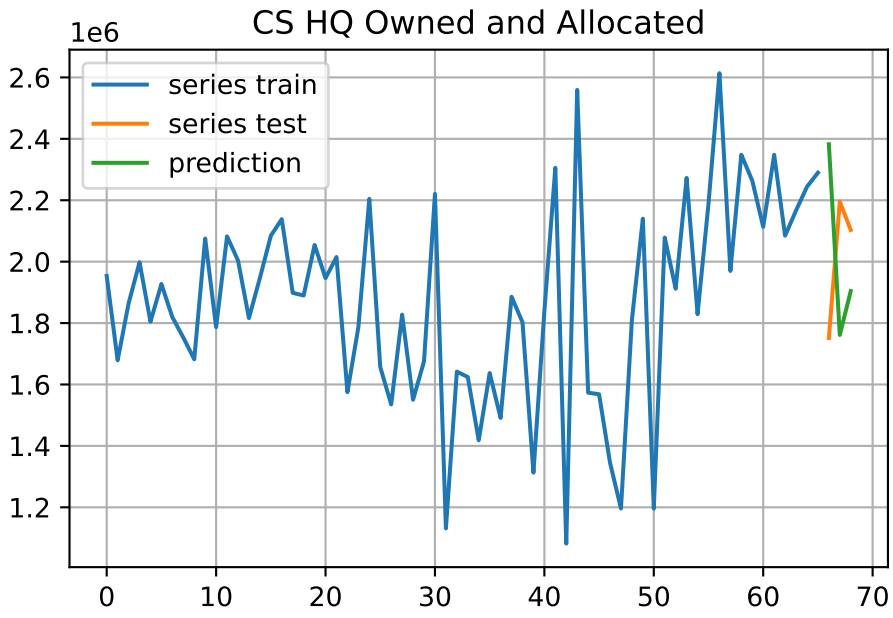


3.1 Red Neuronal

```
nnet=keras.Sequential([
                      keras.Input(shape=(10,)),
                      Dense(11,activation=tf.nn.tanh),
                      Dense(1,activation='linear')
])
nnet.compile(loss='mean_squared_error',optimizer='sgd')
nnet.summary()
```

```
Model: "sequential"

_________________________________________________________________

 Layer (type)                Output Shape              Param #

=================================================================

 dense (Dense)               (None, 11)                121



 dense_1 (Dense)             (None, 1)                 12
```

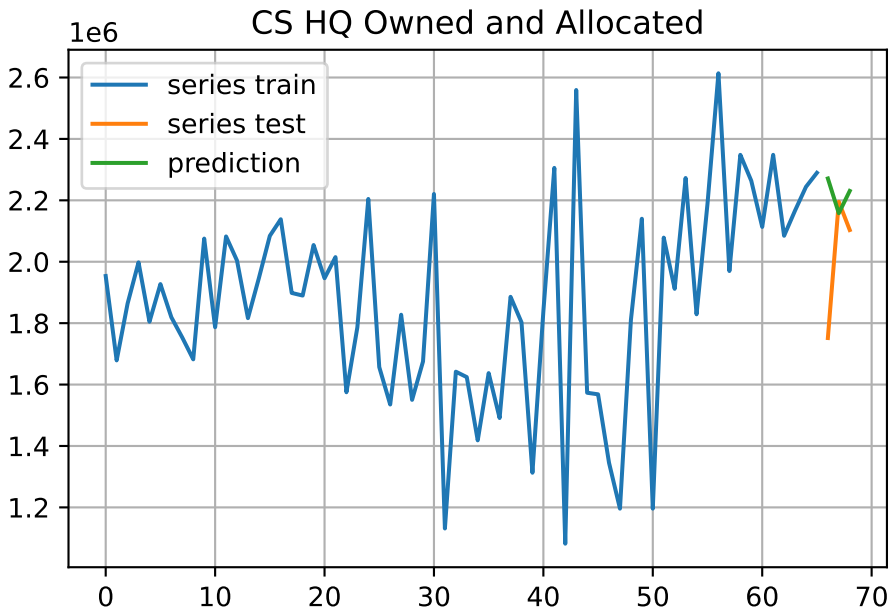
```
=================================================================

Total params: 133

Trainable params: 133

Non-trainable params: 0

_________________________________________________________________
```



3.2 Random Forest

```
model_tree = RandomForestRegressor(n_estimators=1000,
                                   criterion='mse',
                                   max_depth=None,
                                   min_samples_split=2,
                                   min_samples_leaf=1,
                                   max_features='auto',
```

```
                    bootstrap=True,
                    oob_score=False,
                    random_state=0,
                    verbose=1)
```



line_cost month	CS HQ Owned and Allocated
2022-08-01	-0.296661
2022-09-01	0.017025
2022-10-01	-0.060561

3.3 Resultados

Los siguientes resultados muestran las estimaciones de las 6 series de tiempo, comparado para cada modelo (random forest y neural net), para cada mercado (Latin América, Norte América). Cada serie y para cada modelo provee 3 pronosticos. Eso da como resultado que se tienen que evaluar y comparar al menos 72 valores. La forma de comparar es por el error como porcentaje, al igual que el benchmark discutido arriba.

Dos notas previas:

- Porque se necesita hacer varias veces más y que sean modelos similares en la estructura no se optimizaron hiper parámetros

- Porque el modelo es dependiente de valores ya conocidos (pronosticados), si sus datos y estimaciones están mal nuestro modelo y resultado estarán mal.

3.3.1 Norte América

month	CS HQ Owned and Allocated	Contact Center Expense	\
2022-08-01	-0.296661	-0.126735	
2022-09-01	0.017025	-0.194134	
2022-10-01	-0.060561	-0.100452	

month	Contact Center OH	Delivery	GBU Owned and Allocated	Supply Chain
2022-08-01	-0.189576	0.001827	-0.281562	0.097098
2022-09-01	-0.512825	-0.327191	-0.025992	-0.029995
2022-10-01	-0.747165	-0.000289	0.181925	0.085159

month	CS HQ Owned and Allocated	Contact Center Expense	\
2022-08-01	-0.306752	-0.118409	
2022-09-01	0.126551	-0.098854	
2022-10-01	0.042051	0.032315	

month	Contact Center OH	Delivery	GBU Owned and Allocated	Supply Chain
2022-08-01	-0.262581	0.012935	-0.238821	0.116411
2022-09-01	-0.336484	-0.378833	0.032649	0.113745
2022-10-01	-0.802051	-0.053936	0.286099	0.131946

3.3.2 Latin América

month	CS HQ Owned and Allocated	Contact Center Expense	\
2022-08-01	-0.265612	-0.098463	
2022-09-01	0.038745	-0.021356	
2022-10-01	-0.045609	-0.255822	

month	Contact Center OH	Delivery	GBU Owned and Allocated	Supply Chain
2022-08-01	-0.007192	-0.139490	-0.204051	0.153827
2022-09-01	-0.096557	-0.042276	-0.259313	-0.111910
2022-10-01	-0.241665	0.120782	0.052588	0.220457

month	CS HQ Owned and Allocated	Contact Center Expense	\
2022-08-01	-0.155756	0.347851	
2022-09-01	-0.098588	0.123439	
2022-10-01	0.067408	-0.474014	

month	Contact Center OH	Delivery	GBU Owned and Allocated	Supply Chain
2022-08-01	-0.382724	-0.073510	-0.677740	0.205079
2022-09-01	-0.124940	0.147143	-0.784271	-0.295888
2022-10-01	-0.264305	0.240566	-0.014005	-0.063511

Como se puede apreciar, por los errores y las gráficas, algunas de las variables tienen brincos inesperados en t+1. En estimaciones t+2 y t+3 los errores son menores. Lo que me hace pensar que los modelos hacen un buen trabajo generalizando lo que debería ser en base a los valores previos y los otros datos asociados a los costos operativos.

Para algunos tipos de costos parece funcionar bien pero para otros no. En general no supero el benchmark. Los datos tienen un comportamiento errático, lo que ocasiona que los modelos dependientes de datos pasados no funcionen de la mejor manera.

Se necesita probar con otro modelos más sofisticados como LSTM y sencillos como promedios moviles. Además, estudiar a detalle un modelo que no sea únicamente dependiente de su propia serie, conseguir datos que expliquen el fenómeno y por último reducir las irregularidades por errores humanos.

4 Anexos

4.1 Definición de costos

Costo	Explicación
Region Owned Expense	Costos pertenecientes a la region, propios de operaciones involucradas en reparación y asistencia.
Variable Expense	Costos variables de operaciones relacionados a la repación, insumos o asistencia a productos.
Contact Center	Costos variables de asistencia telefónica.
Delivery	Costos variables de reparación y asistencia física.
Supply Chain	Costos variables de la cadena de suministro, insumo de partes, impuestos, logistico e inventario.
Repair OH Expense	Costos fijos o semifijos relacionados empleados de la administración y soporte de las operaciones diarias.
Contact Center OH	Costos fijos o semifijos de empleados administrativos para Contact Center.
Delivery OH	Costos fijos o semifijos de empleados administrativos para el grupo de técnicos e ingenieros.
Supply Chain OH	Costos fijos o semifijos de empleados administrativos para el grupo de cadena de suministro.
Other Warranty Expense	Otros regionales.
Worldwide Owned and Allocated Expense	Costos fijos o semifijos de empleados e inversiones globales que soportan a los tres grupos operativos de CS (Customer Support).
CS HQ	Costos fijos de empleados globales que soportan a los tres grupos operativos, incluyendo administrativos, finanzas y directivos.
CS Investments	Costos fijos de inversiones globales.
GBU Owned and Allocated	Costos de tipo diverso, fijo o variables, que incluye empleados y costos operativos de las unidades globales de negocio (Global Business Unit).
Net Reserve Expense	Reserva neta es la suma de Reserva (Accrual) más Amortización (Amortization). Tipicamente un número positivo.

Costo	Explicación
Accrual for Shipments	Reserva de dinero que la compañia realiza con el motivo de hacer frente a sus obligaciones y pagar a sus empleados y proveedores. Basado en el costo promedio y el porcentaje de fallas esperadas del producto vendido.
Amortization	Amortización de la reservea, siempre un número negativo.