INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Proyecto Estadística Aplicada III

Explorando la distribución del transporte público en la CDMX

Alonso Martinez Cisneros Juan Carlos Sigler Priego Carlos Delgado
Esmeralda Altamirano

Primavera 2022

Índice

1. Propuesta de proyecto

Buscamos resolver la siguiente pregunta clave que pudo haber sido planteada por gobierno o algún particular: *¿Está bien distribuido el sistema de transporte público para servir a las necesidades de la población de la ciudad?*

Buscamos tener una visión histórica de cómo ha ido evolucionando la cobertura del transporte público, para esto tomamos el número total de estaciones de transporte por alcaldía, la facilidad de acceso al transporte más cercano, para lo cual proponemos la distancia promedio de las zonas residenciales a su estación más cercana, y la distancia promedio de las zonas residenciales al zócalo. Con esto último pretendemos responder si alejarse de la zona centro empeora la cobertura sin importar la población. Otro punto al que le damos peso es a la densidad de población ya que lógicamente serían las zonas que más necesitan el transporte. Finalmente es importante destacar que todas las variables anteriores se analizan bajo dado su cambio con el paso de cada año.

Para verificar qué técnicas estadísticas podemos usar analizamos la correlación entre las variables con énfasis en buscar la existencia de una correlación positiva entre la distancia al zócalo de la ciudad y la conectividad del transporte, esto con las variables de distancia al transporte más cercano y el total de estaciones. Para estudiar conectividad como concepto construimos un índice de conectividad para cada delegación, donde buscamos encontrar qué determina los cambios en conectividad por alcaldía. Exploramos si cambios positivos en la conectividad de una zona son producto de un cambio negativo en otra zona.

Finalmente usamos un modelo de regresión lineal para tratar de establecer e interpretar relaciones directas entre las variables de población y año con el índice de conectividad. Para cimentar la correcta interpretación presentamos brevemente pruebas que sustentan que se cumplen los supuestos del modelo.

2. Planteamiento del problema

La Ciudad de México es una de las 10 ciudades más grandes del mundo por población con habitantes al 2021, sin contar la zona metropolitana que incluye zonas del Estado de México e Hidalgo. Una población de este tamaño exige un sistema de transporte público masivo de alta frecuencia, volumen y disponibilidad. El sistema de transporte público unificado en la Ciudad de México es en nuestra opinión uno relativamente bien planeado y accesible. Sin embargo, como personas que no vivimos en la periferia de la zona metropolitana nuestras opiniones pueden estar sesgadas.

El objetivo de esta investigación es cuantificar el nivel de acceso de la población de distintas alcaldías de la ciudad a los diversos medios de transporte público masivo. Como transporte público masivo estamos tomando en cuenta los siguientes servicios de transporte unificado que ofrece la ciudad:

- Metro
- Metrobús
- Tren Ligero
- Cablebús

Elegimos concentrarnos en estos servicios por las siguientes características:

1. Frecuencia. La frecuencia con la que pasan nuevos convoyes debe ser relativamente alta. Por ejemplo, en hora pico pasan convoyes nuevos de metro y metrobús en pocos minutos.
2. Volumen. Nos concentramos en transportes de alto volumen, excluyendo peseros y microbuses.

3. Unificado. Nos concentramos en el sistema de transporte unificado coordinado por el gobierno de la Ciudad de México.

Además de hacer una exploración, el objetivo de esta investigación es determinar que tan bien distribuido está el transporte público en la ciudad. Como habitantes de la CDMX, tenemos la sospecha de que el transporte público está muy centralizado en la zona del centro histórico. Es decir, sospechamos que el sistema de transporte público privilegia a las personas que viven en las delegaciones como Benito Juárez, Cuauhtémoc, etc… que no son necesariamente las delegaciones con las poblaciones más altas.

Estas relaciones las exploraremos mediante diversas técnicas cubiertas en el curso. Primero que nada, procedemos con análisis exploratorio para empezar a ganar intuición sobre el conjunto de datos. Más tarde aplicamos técnicas estadísticas para construir algo como un "índice de conectividad". Exploramos cómo se relaciona este índice con variables de interés como: centralidad medida en distancia a la zona del centro histórico, población, y otros factores.

3. Análisis exploratorio

Hay 16 alcaldías. Cada alcaldía contiene datos de población y conectividad con el transporte público desde el año 1969 hasta el 2021.

Para el análisis que vamos a llevar a cabo recabamos información de diversas fuentes para indicadores de movilidad para las alcaldías de la CDMX y cómo han evolucionado desde principios de la década de los 70.

Tenemos información de 16 alcaldías, las cuales son:

```
##  [1] "Azcapotzalco"        "Coyoacán"            "Gustavo A. Madero"
##  [4] "Iztacalco"           "Iztapalapa"          "Tlalpan"
##  [7] "Tláhuac"             "Xochimilco"          "Benito Juárez"
## [10] "Cuauhtémoc"          "Miguel Hidalgo"      "Venustiano Carranza"
## [13] "Cuajimalpa"          "Magdalena Contreras" "Milpa Alta"
## [16] "Álvaro Obregón"
```

Tenemos información de cada año desde 1969, que es cuando se construye la primera linea del metro, hasta 2021, que es cuando se construyen la líneas del cablebús.

Primero que nada, vemos algunas estadísticas de resumen de los datos.

Tabla 1: Estadística descriptiva del conjunto de datos.

AÑO	ALCALDIA	POBLACION	MEAN_DIST	EST_TOTAL	ZOC_DIST
Min. :1969	Length:848	Min. : 30700	Min. : 140.8	Min. : 0.00	Min. : 2028
1st Qu.:1982	Class :character	1st Qu.: 257079	1st Qu.: 382.5	1st Qu.: 0.00	1st Qu.: 3641
Median :1995	Mode :character	Median : 443768	Median : 942.3	Median : 5.00	Median : 5831
Mean :1995		Mean : 532742	Mean : 2252.4	Mean : 10.94	Mean : 7366
3rd Qu.:2008		3rd Qu.: 639251	3rd Qu.: 3051.8	3rd Qu.: 15.00	3rd Qu.:12048
Max. :2021		Max. :1840000	Max. :12926.4	Max. :118.00	Max. :17697
			NA's :53		NA's :53

3.1. Descripción de las variables de interés

Las variables son las siguientes:

- `AÑO`: El año en el cual están medidas las variables.
- `ALCALDIA`: La demarcación territorial de delegación o Alcaldía al que corresponden los datos.

- `POBLACIÓN`: La población para cada alcaldía en el año dado.
- `MEAN_DIST`: La distancia promedio de todas las zonas marcadas como residenciales en la encuesta de uso de suelo a su estación de transporte público masivo más cercano medida en metros.
- `EST_TOTAL`: Número total de estaciones de transporte público masivo en la alcaldía al año marcado.
- `ZOC_DIST`: Distancia promedio de las zonas residenciales al zócalo de la ciudad.

Las variables fueron construidas a partir de diferentes conjuntos de datos abiertos al público. No encontramos una base de datos que tuviera lista para usarse toda la información que era necesaria para el análisis, menos aún como función del tiempo. En las siguientes secciones ahondamos en algunos detalles técnicos de cómo se obtuvieron, limpiaron, y trabajaron datos faltantes.

3.1.1. Número total de estaciones por delegación

Para encontrar el número de estaciones de transporte público masivo por delegación a un año dado, utilizamos los conjuntos de datos [**metrobús**; STC22; STE21]. En la figura 1 se pueden ver todas las líneas de transporte público consideradas sobre el mapa de la CDMX con división política.

Como se puede ver en la figura, hay razón para sospechar que el transporte público está concentrado al centro y norte del territorio, al menos a primera vista y si no se conoce bien la ciudad. La mayor carencia aparente es al sur de la ciudad. En la figura 16 en el apéndice A presentamos un mapa de la ciudad con divisiones políticas para hacer más fácil referirnos a alcaldías específicas.

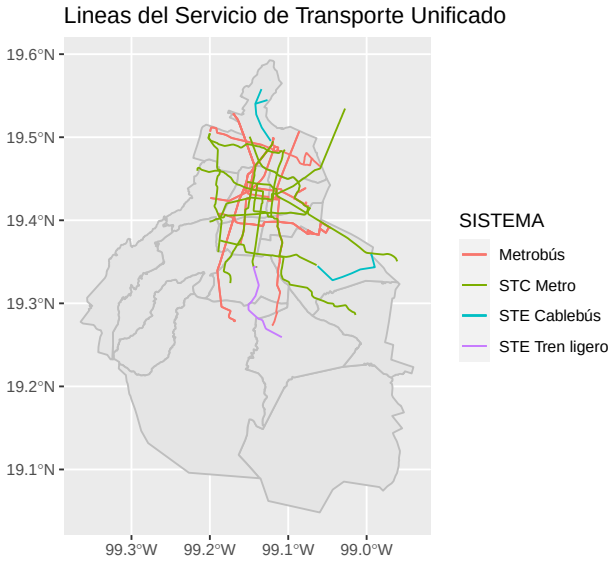


Figura 1

A partir de ahora nos referimos a la "zona centro" como la zona comprendida por las alcaldías: Miguel Hidalgo, Cuauhtémoc, Benito Juárez. Es precisamente esta zona en la que sospechamos está sobre-concentrado el transporte público.

Las carencias más grandes se pueden ver en las alcaldías de Tlalpan, Magdalena Contreras, Xochimilco y Milpa Alta. A comparación de las alcaldías al centro, las alcaldías en el sur tienen pocas estaciones, pocas líneas, y una baja cobertura en general. Más tarde tomamos en consideración la población y otros factores para comparar que tan fácil es acceso de la población de una alcaldía al sistema de transporte unificado. Uno de los factores clave para este análisis es la variable que llamamos `MEAN_DIST`: una métrica que utilizamos para medir que tan fácil es el acceso de una alcaldía al transporte unificado, la cual explicamos con más profundidad a continuación.

3.1.2. Distancia promedio al transporte público

Para calcular una medida de acceso al transporte público nos pareció que tomar solamente la cantidad de estaciones en total contenidas dentro de los límites de una alcaldía sería muy insuficiente. Por ejemplo, alcaldías como Cuajimalpa y Magdalena Contreras que no tienen ninguna

estación estarían efectivamente "desconectadas", pero eso no quiere decir que sus habitantes no tengan manera alguna de transportarse.

Para estimar la "conectividad" tomamos información sobre el uso de suelo de la CDMX publicado por la Secretaría de Desarrollo Urbano y Vivienda [Des21]. Con esta información tomamos la localización de todas las zonas registradas como habitacional o residencial (e.g. habitacional comercial, habitacional multifamiliar) y usando un algoritmo conocido como BallTree calculamos a que distancia, medida con la métrica Haversine, está de la estación de transporte unificado más cercana. Así obtenemos para cada zona residencial una distancia en metros, y después calculamos la media para cada alcaldía para cada año. Utilizamos esta información más tarde para hacer en análisis sobre conectividad por población al que hacíamos referencia.

Decidimos calcularlo de esta manera para tener una mejor idea de cómo es que el transporte está distribuido con respecto a la *población* y dónde vive ésta. Si tomáramos, por ejemplo, el número total de estaciones normalizado por área las alcaldías como Milpa Alta o Tlalpan mostrarían un sesgo considerable dado que son muy grandes en términos de área pero su población es mucho menor a otras alcaldías mucho más pequeñas. Teniendo distancia promedio medida en metros con la métrica Haversine y además la población podemos controlar tanto por el efecto de densidad poblacional como el fenómeno de distribución de la misma. Para dar otro ejemplo, si se tomara la distancia con los extremos de los límites de la alcaldías, veríamos que Cuajimalpa está peor conectado de lo que está en realidad. ¿Por qué? Porque por un extremo tenemos la zona Observatorio y por la otra Desierto de los Leones, que está mucho más lejos de la zona de cobertura del transporte, pero la población ahí es mucho más pequeña que la de la zona Observatorio, por poner un ejemplo.

Vale la pena mencionar que una de las debilidades de este análisis es la falta de información completa. En el conjunto de datos que se utilizó para obtener la distancia promedio no hay ningún registro de las zonas habitacionales para la alcaldía Álvaro Obregón, a pesar de que es una de las más pobladas. Ignoramos la razón de esta falta de datos, pero es razonable pensar que hay otras carencias que no se pueden distinguir a simple vista y que podrían estar sesgando nuestro análisis.

3.1.3. Número total de estaciones & distancia a la zona centro

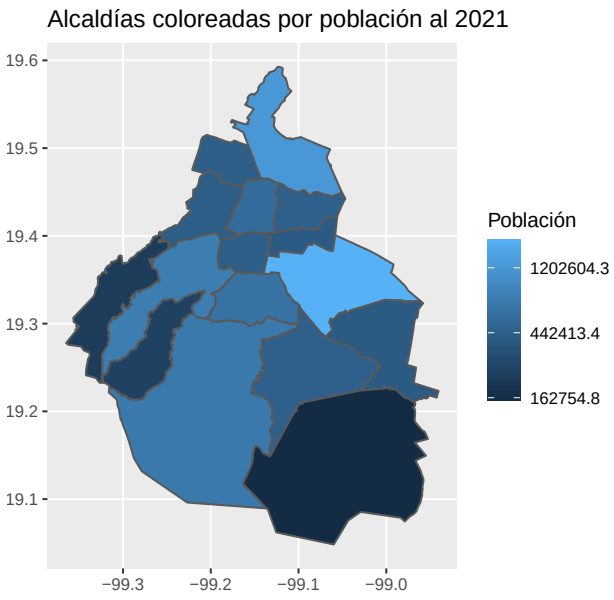


Figura 2

Para complementar nuestro análisis de conectividad, utilizamos otras dos variables calculadas a partir de los conjuntos de datos ya mencionados. La primera de estas variables es el número total de estaciones de transporte público unificado que se encuentran dentro de los límites de una alcaldía dada para algún año fijo. Esto para tomar en cuenta cómo ha evolucionado el sistema de transporte unificado.

La distancia a la zona centro se toma como la distancia promedio en la métrica Haversine de las mismas zonas residenciales, mencionadas en el párrafo anterior, al zócalo de la ciudad. Una vez más, se toma el promedio de estas distancias para la alcaldía y el año correspondiente.

Reconocemos que la elección del zócalo de la ciudad como punto central puede parecer arbitraria. Sin embargo, nos parece justificable puesto que es una de las zonas más antiguas, y por lo tanto, el crecimiento de la zona

metropolitana de la ciudad ha sido radialmente hacia afuera de esta zona. De manera similar, las primeras estaciones de metro y metrobús fueron construidas precisamente para servir a la zona centro.

Con todas las variables a las que hicimos referencia podemos empezar el análisis principal y el objetivo de este trabajo.

3.2. Acceso a transporte con base en la población

Como se mostró antes, el mapa de líneas de transporte público unificado muestra una concentración alta en la zona centro (alcaldías Cuauhtémoc, Miguel Hidalgo y Benito Juárez). Esta concentración sería deseable si éstas fueran las alcaldías más pobladas, ya que justamente una mayor densidad poblacional justifica mayor inversión en el sistema. Con ayuda de la figura 2 podemos ver cómo es la distribución geográfica de la población en la CDMX.

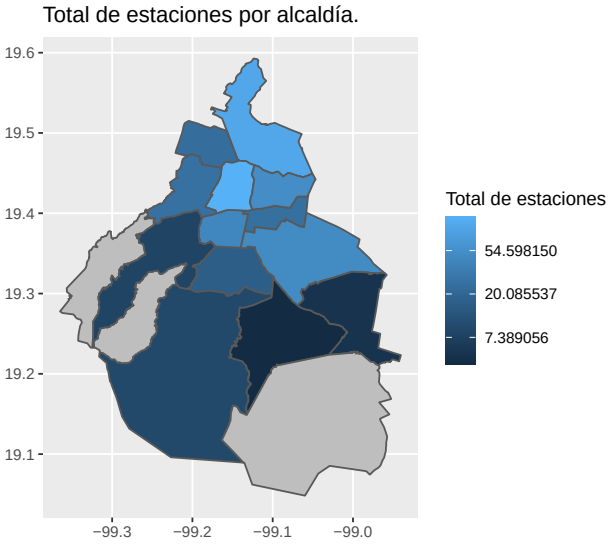


Figura 3

Llama la atención que las alcaldías del centro efectivamente no son las más pobladas. Las dos alcaldías más pobladas son Iztapalapa, Gustavo A. Madero (GAM) y Álvaro Obregón. Tanto Iztapalapa como GAM están en la periferia de la ciudad, y ninguna de las tres más pobladas está en la "zona centro".

En la figura 3 vemos las alcaldías coloreadas dependiendo de cuantas estaciones de transporte público al año 2021 tienen en total, y en la figura 4 podemos ver el total de estaciones por alcaldía como función del tiempo.

Las alcaldías con el mayor número total de estaciones de transporte público al año 2021 son Cuauhtémoc, GAM y Venustiano Carranza en ese orden. De la lista de alcaldías más pobladas solo coinciden Gustavo A. Madero. Notablemente, Iztapalapa parece tener un déficit de transporte público al ser la alcaldía más poblada por un margen alto, con más de 1 millón de habitantes, pero siendo la cuarta con más estaciones de transporte público. También llama la atención que hay tres alcaldías que no tienen una sola estación de transporte público: Cuajimalpa, Magdalena Contreras y Milpa Alta. En el caso de Milpa Alta tiene sentido dada la baja densidad poblacional, pero en Cuajimalpa no solo hay áreas densamente pobladas, sino que hay áreas de suma importancia comercial, como la zona de Santa Fe.

Hasta el momento hemos tomado la información en el punto de tiempo más reciente al que tenemos acceso: al año 2021. Para hacer un análisis más robusto tomamos en cuenta el componente temporal y estudiamos cómo ha cambiado la "conectividad" de diversas alcaldías con el paso del tiempo.

En la figura 13b se puede ver un *boxplot* que ayuda a entender la evolución de la conectividad como función del tiempo. En ella comparamos las distancias promedio a la estación de transporte público más cercano por alcaldía, donde cada observación corresponde a un año. Dado que esta distancia es estrictamente decreciente (no se ha dado el caso de que se elimine por completo una estación permanentemente), la dispersión de los datos nos dice cómo se ha ido reduciendo esa distancia desde que se creó la primera línea del metro hasta la actualidad. En la figura 4 nos ayuda a ver específicamente qué años marcan la diferencia y así podemos estudiar qué cambio mejora o empeora la conectividad de una alcaldía

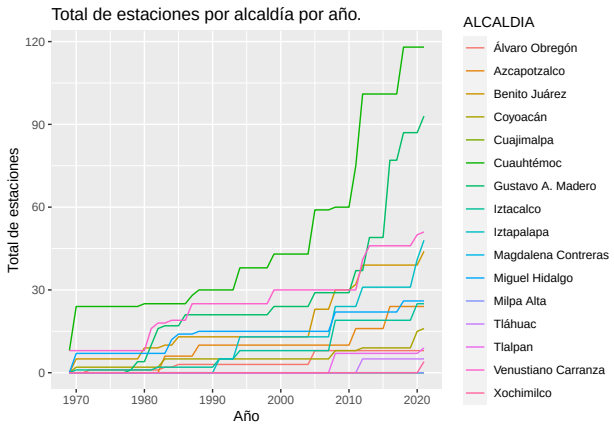


Figura 4

Podemos ver que las alcaldías de Cuauhtémoc, Benito Juárez y Venustiano Carranza tienen las menores varianzas en distancia media y además las más pequeñas. Estas alcaldías son precisamente la que definimos como "zona centro" desde el inicio. De esta observación confirmamos que la zona centro siempre ha estado muy bien conectada, porque el sistema de transporte unificado fue construido pensando en servir específicamente a esta zona. Además, se puede notar que su distancia promedio a la estación de transporte más cercana sigue siendo muy baja en comparación a otras alcaldías, incluso las más pobladas como Iztapalapa.

Por otro lado, Tláhuac, Cuajimalpa y Milpa Alta son los de mayor distancia y variación. En el caso de Tláhuac por ejemplo, siendo una de las alcaldías más al sur, lo que interpretamos es que su distancia promedio a las primeras estaciones era excesivamente alta y fue disminuyendo a medida que mejoró la cobertura. En el caso de Cuajimalpa la distancia disminuyó dramáticamente pero al día de hoy, sigue siendo la alcaldía "peor conectada" por distancia.

Otra cosa que podemos observar es que la línea en la caja que marca la media está en todos los casos mucho más cerca del extremo izquierdo de la caja. Lo cual nos quiere decir que los datos están sesgados, y que la mayoría está más cerca del lado de "distancia baja". En otras palabras, la distancia promedio mejoró rápidamente, lo cual sugiere que el sistema de transporte unificado evolucionó rápidamente para cubrir gran parte de la zona metropolitana.

En la figura 4 consideramos el número total de estaciones en la alcaldía como función del tiempo. Aquí podemos ver que el número de estaciones en las alcadías de la zona centro excede vastamente el de las alcaldías más periféricas, como Iztapalapa. Analizando la variabilidad mediante el ancho de la caja podemos ver también que, por ejemplo, en la alcaldía Cuauhtémoc y GAM se han construido muchas estaciones con el paso de los años. Lo cual nos da pistas, por ejemplo, en el caso de Cuauhtémoc que no solo comenzaron estando muy bien conectadas, la inversión ha continuado más y más a pesar de que era buena desde un inicio. El número total de estaciones en Cuauhtémoc ha llegado a casi 120, mientras que en la mayoría no se exceden las 50.

Otra cosa que llama la atención es el caso de Benito Juárez. El número total de estaciones no ha crecido tan dramáticamente como en las otras alcaldías de la zona centro, pero recordando su distancia promedio al transporte es una de las alcaldías mejor conectadas. Esto nos indica que a pesar de que no se han hecho muchas estaciones nuevas en sus límites territoriales, las que se hicieron han estado en la zona circundante y mejoraron su conectividad. Si recordamos el mapa, la alcaldía Benito Juárez colinda con Cuauhtémoc al norte, que es una de las alcaldías que ha recibido mayor inversión, como se puede ver por el elevado número de estaciones del sistema de transporte unificado.

Esto quiere decir que la inversión beneficia indirectamente a las alcaldías vecinas de la alcaldía en la que se construyen nuevas lineas o estaciones, pero con un detalle: solo si no tenían sus propias estaciones. Análogamente, construir nuevas estaciones en Benito Juárez no beneficia a Cuauhtémoc en términos de reducir su distancia promedio al sistema unificado, porque ya tenía estaciones propias mucho más cercanas.

3.2.1. Análisis de Correlación

Si bien, hasta ahora nos hemos servido de interpretar diversas gráficas para tomar intuición, si queremos cuantificar qué tan notorio es el efecto de inversión privilegiada en la zona centro, tenemos que servirnos de otras técnicas estadísticas. Por ejemplo, si nuestra hipótesis tiene evidencia favorable esperaríamos observar una correlación positiva entre distancia al zócalo de la ciudad y la conectividad medida como distancia promedio al transporte más cercano y número total de estaciones. Además, si pretendemos utilizar alguna técnica estadística como regresión lineal es importante que haya correlación entre las variables. De otra forma no podremos explicar a una utilizando el resto. En la figura 5 vemos un diagrama de correlación para las variables estudiadas.

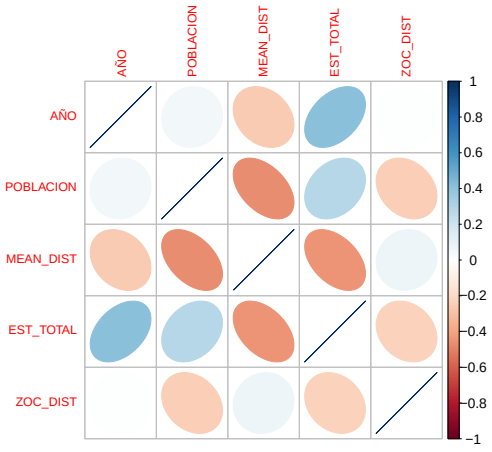


Figura 5: Gráfica de correlación entre las variables del conjunto de datos.

Efectivamente se cumple que la correlación de distancia al zócalo con distancia al transporte más cercano es positiva. Es decir, entre más se aleja la zona habitacional del zócalo, más se aleja de la zona de cobertura del sistema de transporte unificado. También se puede apreciar este fenómeno en la correlación negativa entre distancia al zócalo con el número total de estaciones. Es decir, entre más lejos está la alcaldía del zócalo menor es el número total de estaciones a las que se tiene acceso. Las correlaciones son aparentemente débiles, pero notables. Sospechamos que la correlación se hace más fuerte a medida que se va hacia atrás en el tiempo cuando había menos estaciones en total. El corolario es que esta conectividad si ha mejorado desde que se empezó a construir la primera linea de metro hasta la actualidad. Más tarde profundizamos en esta posibilidad y analizamos la conectividad medida mediante un constructo como la relación de este con otras variables segmentando el conjunto de datos año por año. Por ahora el análisis general parece prometedor.

La matriz explícita de correlaciones se puede encontrar en la tabla 2. Como se puede ver, las correlaciones no exceden el 0.5 en valor absoluto, pero tampoco son insignificantes. Más tarde probaremos con formalidad si está o no justificado un análisis mediante otras técnicas dadas estas correlaciones observadas.

Tabla 2: Matriz de correlación entre las variables de interés.

	AÑO	POBLACION	MEAN_DIST	EST_TOTAL	ZOC_DIST
AÑO	1.000	0.050	-0.253	0.411	0.000
POBLACION	0.050	1.000	-0.461	0.290	-0.249
MEAN_DIST	-0.253	-0.461	1.000	-0.443	0.077
EST_TOTAL	0.411	0.290	-0.443	1.000	-0.222
ZOC_DIST	0.000	-0.249	0.077	-0.222	1.000

En la figura 6 vemos la a distancia promedio al sistema de transporte por alcaldía con un mapa. Si observáramos un color menos brillante a medida que nos alejamos del centro de la ciudad tendríamos una pista de que nuestra hipótesis es cierta y el sistema privilegia al centro de la ciudad. Como medida de visualización está bien, pero hay varios problemas con ella como método formal. Por ejemplo, que algunas alcaldías son muy "largas" y sus puntos más cercanos y más lejanos al centro de la ciudad serán coloreados del mismo color a pesar de que no tienen la misma conectividad. El mejor ejemplo de este caso es Álvaro Obregón. Su zona norte y oriente están bien conectadas: cerca de Tacubaya y prolongación paseo de la Reforma respectivamente. Por otro lado, las zonas como Los Dinamos y Las Águilas están muy lejos del resto del sistema.

Este fenómeno se repite con otras alcaldías "largas" como Magdalena Contreras. Difícilmente se pueden comparar zonas en la misma delegación como Santa Teresa, que esta densamente

poblada, con zonas boscosas como el Ajusco.

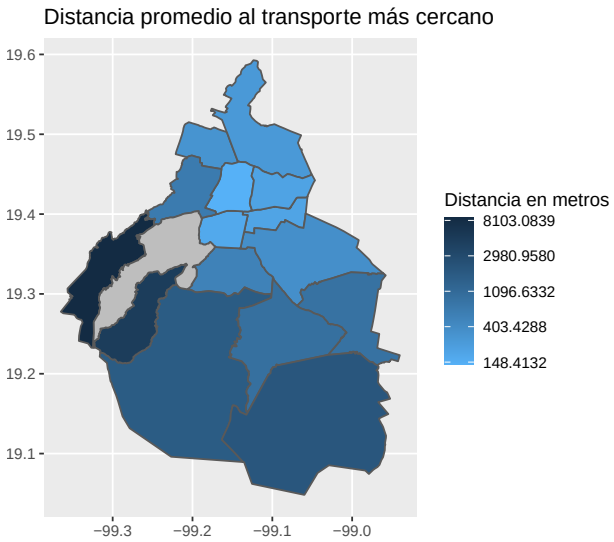


Figura 6

Los mapas hasta ahora solo han mostrado la información referente al año más actual: 2021. En la figura 7 vemos cómo ha evolucionado esta distancia promedio a medida que avanza el tiempo.

La caída pronunciada de las líneas correspondientes a varias alcaldías coincide con la apertura de la línea morada que va de Pantitlán a Santa Martha Acatitla. Entre las alcaldías que ven el cambio más dramático están Milpa Alta y Tláhuac. Sospechamos que esto se debe a que esta línea está al norte de Iztapalapa y antes de eso el transporte estaba aún más concentrado en el centro, lo cual hizo a esa línea relativamente lejana, la opción más cercana.

En el caso de Iztapalapa, podemos ver que antes de 1990 era una de las alcaldías con la distancia promedio al transporte más altas, siendo comparable con Cuajimalpa, Magdalena Contreras y Milpa Alta. Sin embargo cuando se construyó esta línea que está en la frontera norte entre Iztapalapa y los municipio de Ciudad Nezahualcóyotl redujo esta métrica dramáticamente hasta niveles comparables con las alcaldías más céntricas como Venustiano Carranza, Benito Juárez e Iztacalco.

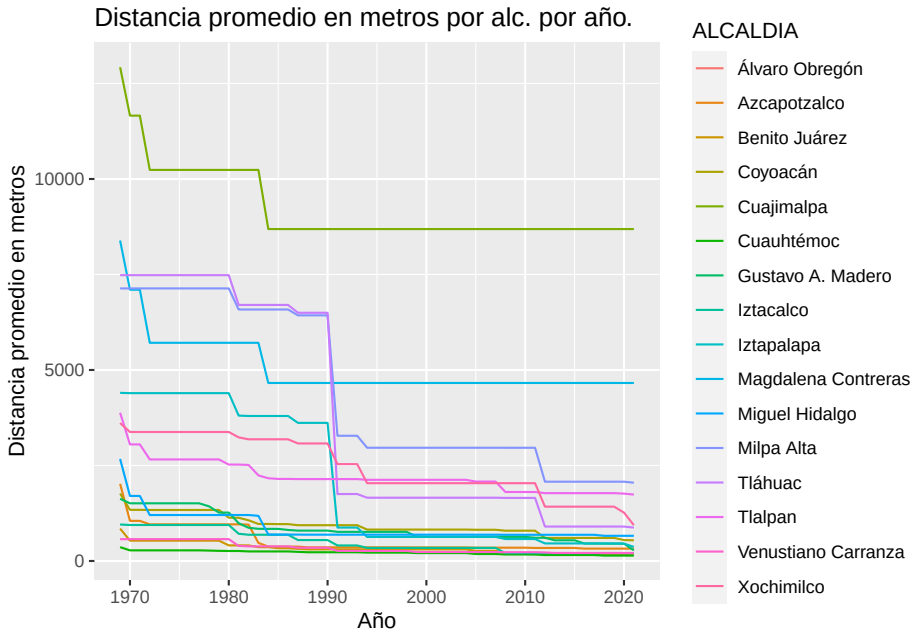


Figura 7

4. Análisis de Componentes Principales

Para identificar que variables dominan y tienen mayor efecto en la varianza hacemos un análisis de componentes principales. Lo primero que notamos es que ninguna componente es

verdaderamente dominante. Podríamos decir que la variabilidad se explica bien hasta la cuarta componente que tenemos el 91 % de ella. La salida de `princomp` se puede ver abajo.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Desviación estándar     1.434476 1.0505992 0.9382218 0.7185053 0.66483939
## Proporción de la varianza 0.411544 0.2207517 0.1760520 0.1032500 0.08840228
## Proporción acumulada  0.411544 0.6322957 0.8083478 0.9115977 1.00000000
##
## Correlación entre variables y las componentes:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## AÑO        0.363  0.643  0.327  0.581  0.105
## POBLACION  0.462 -0.434 -0.400  0.496 -0.438
## MEAN_DIST -0.536          0.450  0.239 -0.672
## EST_TOTAL  0.547  0.195  0.231 -0.597 -0.503
## ZOC_DIST  -0.262  0.599 -0.691        -0.303
```

En términos de interpretación, la primera componente contrasta principalmente la distancia promedio con la población y la cantidad de estaciones. Esto nos diría que las delegaciones más pobladas y con más estaciones son las que tienen menor distancia promedio. Esto es natural porque al haber más gente, hay más zonas residenciales y con más estaciones, hay menos distancia. También indica que la población y la cantidad de estaciones han aumentado con el paso de los años, así como que una gran distancia al zócalo y una baja distancia están asociada a menos población y menos estaciones.

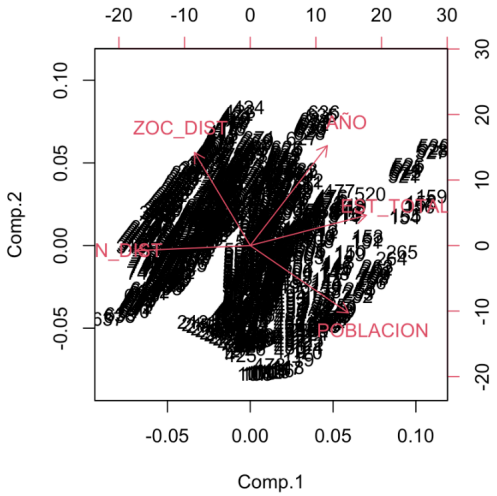


Figura 8: Dispersión según las primeras dos componentes principales

La segunda componente contrasta principalmente la población con la distancia al zócalo, y el paso del tiempo. A mayor distancia al zócalo a través del tiempo, se asocia menos población, quizá indicando que la población se ha concentrado cerca de ahí.

La tercera componente contrasta la distancia al zócalo y la población con la distancia promedio y la cantidad de estaciones. Esta componente está directamente relacionada a nuestra hipótesis. Según esta componente, esta hipótesis podría explicar cerca del 20 % de la situación.

Al hacer un *biplot*, vemos que en las primeras dos componentes, la cantidad de estaciones estaría inversamente correlacionada a la distancia promedio, y que la población está inversamente correlacionada a la cantidad de población, dos aseveraciones de las que ya habíamos hablado antes.

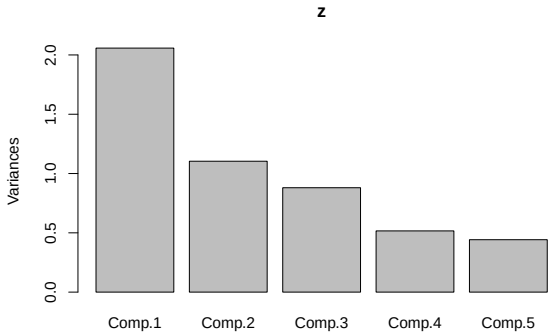


Figura 9: Scree plot para componentes principales del conjunto de datos

5. Construcción de un índice de conectividad

Aquí construimos un índice de conectividad basado en los datos del año 2021. El índice lo construimos por medio del análisis factorial. Las variables utilizadas serán la distancia promedio a las estaciones y cantidad de estaciones en la alcaldía. La prueba de esfericidad de Bartlett indica que las correlaciones son significativas ya que aplicarla da un valor-p de 2.47×10^{-22} y la prueba Kaiser–Meyer–Olkin (KMO) indica una adecuación medianamente regular. El valor es de 0.5, que es apenas suficiente para justificar el uso de esta técnica.

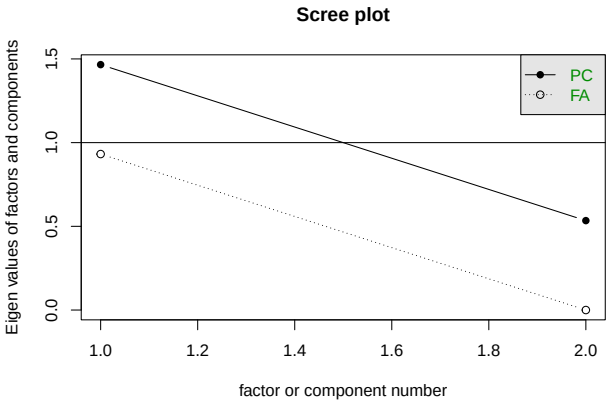


Figura 10: Scree plot comparando componentes principales y análisis de factores

El gráfico de sedimentación (*scree plot* en inglés) en la figura 10 sugiere que un factor es suficiente en este caso.

Queremos combinar la información que dan el número total de estaciones y la distancia promedio a ellas para aproximar una variable intangible: qué tan conectada está la alcaldía al resto del sistema de transporte público unificado. Para esto utilizamos análisis de factores para construir un índice de conectividad. En la tabla 17 (en el apéndice A) se puede encontrar la salida completa de la aplicación de esta función al conjunto de datos. El diagrama que explica cómo está construido el índice y la participación de cada factor que lo compone se puede ver en la figura 11.

Al ver el modelo generado, vemos que el factor o constructo da importancias comparables a la cantidad total de estaciones de transporte en la alcaldía como su distancia promedio a ellas. Un valor muy alto del índice indicaría que hay muy buen acceso en cuanto a número de estaciones, las cuales están bien distribuidas en el territorio lo cual baja la distancia promedio a ellas. Un índice bajo indica que no solo no hay muchas o ninguna estación en el territorio, las más cercanas en otras alcaldías están relativamente lejos.

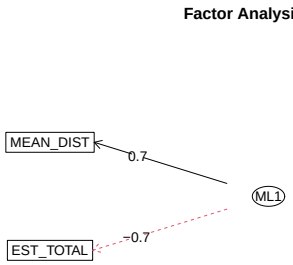


Figura 11: Diagrama de construcción del índice de conectividad

Construimos ahora el índice para cada alcaldía en 2021, el cual se puede encontrar en la tabla 3. El índice de conectividad, hasta ahora, confirma nuestras sospechas. Las alcaldías que muestran mayor accesibilidad al sistema de transporte unificado son las que definimos como zona centro o sus alcaldías colindantes. La más conectada es Cuauhtémoc, lo cual es de esperarse puesto que no solo es la que tiene la mayor cantidad de estaciones por mucho, sino que también es la que está más cerca en promedio a estas estaciones, como consecuencia de su tamaño reducido. En el otro extremo está Cuajimalpa que no tiene ni una sola estación y además está lejos del resto del sistema. De hecho, las alcaldías con los 3 peores índices de conectividad son los que no tienen una sola estación del sistema de transporte unificado.

En la figura 12 podemos ver el mapa de la CDMX con la alcaldías coloreadas de acuerdo a su índice de conectividad. Una vez más, el mapa ayuda a identificar si hay un patrón espacial en los datos. En este caso podemos ver con claridad que a medida que nos alejamos de la alcaldía Cuauhtémoc en cualquier dirección la conectividad baja proporcionalmente a que tan lejos se está. La excepción notable es Gustavo A. Madero, que muestra un índice de conexión cercano

	score
Cuauhtémoc	1.000
Gustavo A. Madero	0.891
Venustiano Carranza	0.725
Iztapalapa	0.704
Benito Juárez	0.699
Iztacalco	0.621
Azcapotzalco	0.609
Miguel Hidalgo	0.597
Coyoacán	0.564
Tláhuac	0.499
Xochimilco	0.491
Tlalpan	0.462
Milpa Alta	0.407
Magdalena Contreras	0.247
Cuajimalpa	0.000

Tabla 3: Índice de conectividad calculado al 2021

en su tonalidad. Una vez más, en el sur de la ciudad se nota la conectividad mucho menor. Más tarde utilizaremos un modelo lineal para analizar si la baja población de estas alcaldías explica adecuadamente este aparente descuido y rezago.

Otro caso a destacarse es el de Iztapalapa. Iztapalapa no es una alcaldía céntrica pero presenta un índice de conectividad mayor al de otras alcaldías que podemos identificar como más céntricas o más afluyentes como Benito Juárez, aunque la diferencia en términos numéricos es apenas destacable. Lo que hace el caso de Iztapalapa llamativo es que a diferencia de Benito Juárez, Iztapalapa es muy grande. El índice alto da pistas de que a pesar de su extensión territorial, el sistema de transporte público ha logrado dar buena cobertura tanto en cantidad de estaciones en total como en la forma en la que estas están distribuidas en el territorio.

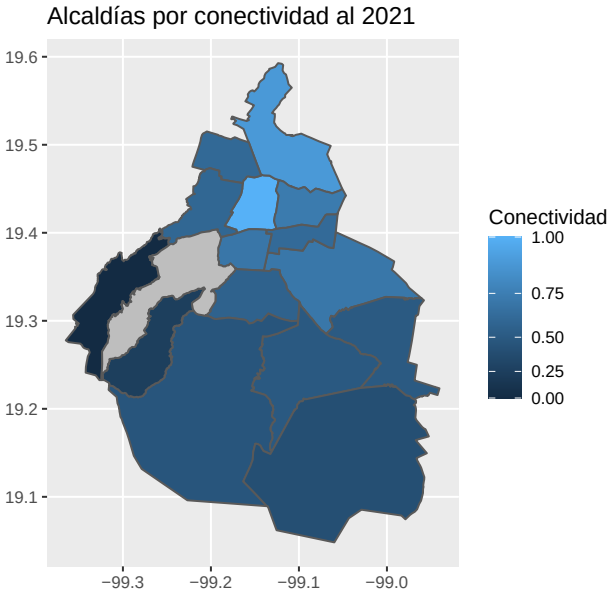
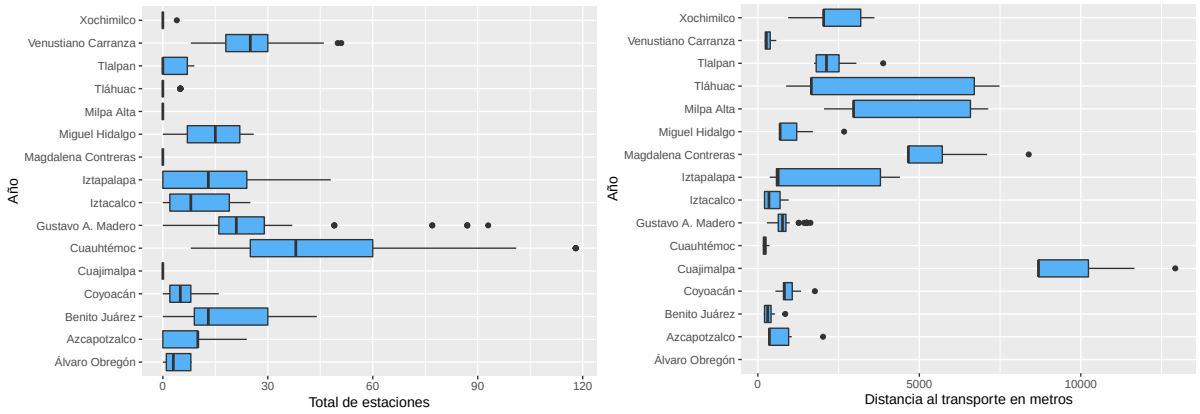


Figura 12: Mapa de las alcaldías por conectividad al 2021

Entre estos dos caso destacables: Gustavo A. Madero e Iztapalapa hay un factor común: la construcción reciente del cablebús. En GAM por ejemplo, la "peninsula" del norte que está rodeada por el Estado de México no había ninguna estación cercana del sistema unificado, toda esa área aumentaba la distancia promedio. En Iztapalapa el cablebús une el centro con el extremo oriente, conectando dos líneas de metro. Si ponemos atención a la figura 1 que muestra las líneas de transporte se puede ver con más claridad el efecto positivo que ha tenido el cablebús. Un análisis mucho más extenso de urbanismo también mencionaría los otros beneficios que tiene como conectar a la población en terreno topográficamente difícil al resto del sistema unificado.

5.1. Análisis del índice de conectividad por año

El mapa nos ayuda a identificar patrones espaciales, pero para notar patrones temporales utilizamos *boxplots*, también conocidas como gráficas de caja y bigotes. Lo que ellas nos permiten es ver mediante su largo cómo es que ha cambiado la variable a través del tiempo. Por ejemplo un *boxplot* muy largo indica mucha variación, o en otras palabras, si nos concentramos en la figura 13a, en el caso de Cuauhtémoc vemos que su caja es de las más largas. De esto inferimos que es donde más se han construido estaciones, ya que pasó de tener relativamente pocas a tener la mayor cantidad. Cuajimalpa por ejemplo tiene un punto en vez de caja porque en todos los años ha tenido la misma cantidad de estaciones: ninguna.



(a) Distribución del total de estaciones por alcaldía a través del tiempo

(b) Distribución de la distancia a la estación de transporte más cercana en metros a través del tiempo

Figura 13: Boxplots para total de estaciones de transporte público y la distancia promedio a éstas por alcaldía.

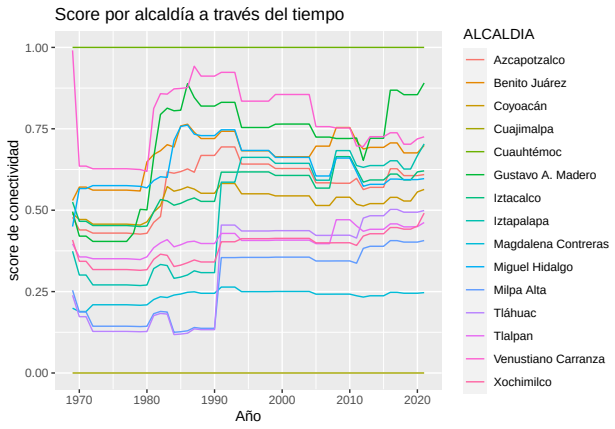


Figura 14: Mapa de las alcaldías por conectividad al 2021

Otra ventaja de utilizar boxplots para analizar la dinámica es que tanto el total de estaciones como la distancia a ellas son funciones monótonas: el número de estaciones no se disminuye porque no se destruyen líneas (exceptuando el caso extraordinario de la línea dorada) y la distancia promedio a estaciones no aumenta por la misma razón. Entonces, el sesgo de los datos nos ayuda a entender en qué punto de tiempo se construyeron las estaciones que conllevan a los cambios en las variables. Que la línea que indica la media de los datos esté sesgada a la izquierda nos indica que en la mayoría de observaciones (es decir años), por ejemplo, el total de estaciones ha sido bajo y que este aumentó recientemente. Un sesgo a la derecha indica lo contrario: que el número de estaciones creció relativamente temprano en esta ventana de tiempo que estamos estudiando, y la mayoría de los años desde 1969 a la fecha los ha experimentado con esa cantidad alta de estaciones.

En ejemplos concretos, en la figura 13b el extremo sesgo que se puede ver en la caja de Cuajimalpa nos indica que su distancia promedio al transporte mejoró apenas recientemente, y que en la ventana de tiempo estudiada la mayoría del transporte se constuyó lejos de Cuajimalpa.

Lo mismo para Iztapalapa, que como ya discutimos mejoró mucho gracias a la construcción de la línea morada y del cablebús. La linea morada se construyó en 1991 y el cablebús en 2021, ambos en la segunda mitad del periodo de tiempo sobre el cual tenemos datos.

Para tener mejor idea sobre en qué años específicamente se dieron los cambios que dieron lugar a los cambios en la conectividad presentamos en la figura 14 la gráfica del índice de conectividad con respecto al tiempo. En ella vemos una vez más los cambios drásticos que ya habíamos comentado respecto a Iztapalapa, Milpa Alta y Tláhuac: el extremo oriente de la ciudad. Damos crédito a esta mejora dramática en conectividad a la linea morada del metro que mejora muchísimo la distancia promedio a la estación de transporte más cercana.

Otra cosa interesante que podemos notar es que, a pesar de que tanto el número total de estaciones como la distancia promedio a ellas son monótonas el índice si sube y baja. Esto se debe a que este score de conectividad empeora cuando el de otras alcaldías mejora, como se puede ver por ejemplo con Azcapotzalco a principios de la década de 1990.

6. Regresión lineal

Para interpretar el papel que juega la población en el incremento de conectividad de una alcaldía, utilizamos un modelo lineal por sus propiedades de interpretación. Estimamos el siguiente modelo:

$$\text{score}_i = \log(\text{AÑO}_i) + \log(\text{POBLACION}_i) + \varepsilon_i, \tag{1}$$

al conjunto de datos que de todas las alcaldías disponibles, salvo Álvaro Obregón, ya que no se pudo calcular el índice por la falta de datos, ni Cuauhtémoc ni Cuajimalpa porque su índice es siempre 1 y 0 respectivamente lo cual causa problemas en la estimación.

Para poder utilizar el modelo primero verificamos que se cumplan sus supuestos: homocedasticidad, no autocorrelación y rango completo. Para eso usamos las pruebas de White, Breusch–Godfrey y el valor inflacionario de la varianza que prueban por cada supuesto respectivamente. Los resultados de aplicar dichas pruebas se pueden encontrar en la siguiente tabla:

Prueba	valor-p o VIF
White	0.000
Breusch–Godfrey	0.000
máx VIF	1.024

Para las pruebas de White y Breusch–Godfrey la hipótesis nula es homocedasticidad y no autocorrelación respectivamente. El VIF (valor inflacionario de la varianza) se usa como regla de decisión, si está por debajo de 30 no hay problemas graves de multicolinealidad. Como se puede ver, los valores-p son ceros numéricos, por lo que no se rechaza la hipótesis nula y se cumplen tanto homocedasticidad como no autocorrelación. Por el otro lado, como el valor VIF más grande es apenas mayor a 1, no tenemos ningún problema de multicolinealidad. En resumen: se puede aplicar el modelo y usar los resultados para dar conclusiones estadísticas.

Los coeficientes estimados se pueden encontrar en la tabla 4.

Se puede ver que los coeficientes asociados a cada variable son altamente significativos. Tanto el año como la población de una alcaldía se relacionan positivamente con su nivel de conectividad, pero el coeficiente asociado al año es mucho mayor. Esto nos indica que, sí el gobierno continúa desarrollando transporte público al mismo paso, se espera que la conectividad de cada alcaldía mejore notablemente, mientras que la población casi no explica el aumento en la oferta de transporte público para una alcaldía dada.

Finalmente llevamos a cabo este mismo análisis, pero ajustando este modelo sobre el conjunto de datos restringido a una sola alcaldía para cada alcaldía. La tabla que contiene todos los

Tabla 4: Estimación de parámetros para modelo lineal descrito.

	Dependent variable:
	log(score)
log(AÑO)	3.617*** (0.484)
log(POBLACION)	0.098*** (0.005)
Constant	−27.672*** (3.674)
Observations	689
R^2	0.446
Adjusted R^2	0.445
Residual Std. Error	0.096 (df = 686)
F Statistic	276.483*** (df = 2; 686)
Note:	*p<0.1; **p<0.05; ***p<0.01

coeficientes estimados se puede encontrar en los anexos, por ahora solo mencionamos ejemplos notables. Entre estos destacan Milpa Alta y GAM.

Tabla 5: Coeficientes del modelo estimado para GAM y Milpa Alta.

ALCALDIA	term	estimate	std.error	p.value	is.signif
Gustavo A. Madero	(Intercept)	-88.973	13.503	0.000	signif
Gustavo A. Madero	log(AÑO)	11.378	1.668	0.000	signif
Gustavo A. Madero	log(POBLACION)	0.426	0.166	0.013	signif
Milpa Alta	(Intercept)	-49.473	50.868	0.335	no signif
Milpa Alta	log(AÑO)	6.515	6.759	0.340	no signif
Milpa Alta	log(POBLACION)	0.049	0.112	0.663	no signif

Para GAM todos los coeficientes son altamente significativos, y el coeficiente asociado a la población es mucho más grande comparado al modelo de todas las alcaldías en conjunto. Por el otro lado, para Milpa Alta ningún coeficiente es significativo al 10 %, lo cual da pistas de que lo que pase en Milpa Alta es de poco interés para las autoridades encargadas de planear y construir transporte público, lo cual no sorprende dada su baja densidad poblacional.

Usando técnicas de regresión lineal podríamos ajustar este modelo a diferentes subconjuntos de datos: las alcaldías del centro y el resto. Así podríamos analizar si los coeficientes son significativamente diferentes. Sin embargo, en interés de la brevedad no incluimos este análisis final.

7. Interpretación, conclusiones, etc...

El punto clave de nuestro análisis fue ver si en efecto la cobertura del transporte público en la ciudad está más enfocada en la zona centro o si tiene está mejor explicada por la densidad de población que puede hacer uso de este. Esto último fue de vital importancia para evitar resultados sesgados, a diferencia de si se hubiera tomado exclusivamente la magnitud del área geográfica o la población total de cada alcaldía sin importar la dinámica de la zona.

Luego, los resultados arrojados de todas las técnicas y modelos aplicados se pueden dividir en dos, los que vienen de un análisis del “presente”, es decir, de como está el transporte público al 2021, y los del “pasado”, donde nos concentramos en su evolución. La razón de ver de esta

forma los resultados es porque la respuesta a nuestra hipótesis puede no ser la misma si no consideráramos el panorama completo.

Por un lado, tenemos que a pesar de que las primeras líneas del transporte público si tuvieron preferencia en la zona centro de la ciudad nuestros análisis arrojaron como resultado que con el paso del tiempo la cobertura de la conectividad ha ido mejorando significativamente en las zonas de la periferia que por esta misma condición tienden a estar más densamente pobladas. Y como predicción, los datos apuntan a que el crecimiento tanto de nuevas rutas de transporte como de estaciones provocará que la mayor parte de la ciudad este conectada de forma aún más eficiente. También resulta claro que la inversión futura probablemente debería hacerse fuera de la zona centro y específicamente la alcaldía Cuauhtémoc para mejorar la conectividad del resto de la ciudad por igual. Por otro lado, pudimos destacar con los datos a los que tenemos acceso es que sigue habiendo zonas que están bastante distanciadas de tener una buena conectividad en comparación con las céntricas, como pudimos ver en el caso de Cuajimalpa, lo cual podría abrir una discusión de porqué se han dejado de lado ciertas zonas periféricas y otras no.

Ahora, las interpretaciones que obtuvimos salen como respuestas a los cambios notorios observados, para empezar que el transporte público en sus primeros años estuviera en la "zona centro" es debido a que la ciudad se ha ido ampliando al hacer parte de ella pueblos que se encontraban en los límites, un ejemplo de esto es la zona sur, como Magdalena Contreras o también Cuajimalpa. No debemos olvidar que la zona Santa Fe, tan icónica como es al día de hoy, es un desarrollo relativamente reciente en la historia de la ciudad. Otro aspecto que no podemos explorar por no tener acceso a los datos es la dificultad que viene de la topografía de ciertos lugares. Por ejemplo sería muy difícil construir metro en Cuajimalpa dado el tipo de suelo y el cambio en elevación.

Finalmente, creemos que este análisis se podría ampliar y mejorar si tuviéramos acceso a más datos, porque el hecho de que exista tan poca información de una zona importante dado su tamaño, ubicación y población, como Álvaro Obregón, limita los resultados e interpretaciones en cierto grado. Y también quedan dudas de porque no se han desarrollado nuevas rutas de transporte en lugares que tienen una densidad de población alta, tal vez analizando datos no solo de las zonas residenciales, sino también de las zonas comerciales, es decir, donde hay más trabajos, y viendo su relación encontraríamos una respuesta a si hay mas variables que afectan la decisión de invertir en el transporte público o si el gobierno ha prestado poca atención a ciertas zonas.

Referencias

[STC22] Secretaría de Transporte Colectivo (STC). *Ubicación de líneas y estaciones del Sistema de Transporte Colectivo Metro*. Portal de Datos Abiertos CDMX, ene. de 2022. URL: https://datos.cdmx.gob.mx/dataset/lineas-y-estaciones-del-metro.

[STCMB] Secretaría de Transporte Colectivo (STC). *Ubicación de líneas y estaciones del Metrobús*. Portal de Datos Abiertos CDMX, ene. de 2021. URL: https://datos.cdmx.gob.mx/dataset/geolocalizacion-metrobus.

[STE21] Servicio de Transportes Eléctricos (STE). *Ubicación de líneas y estaciones/paradas del Servicio de Transportes Eléctricos*. Portal de Datos Abiertos CDMX, oct. de 2021. URL: https://datos.cdmx.gob.mx/dataset/geolocalizacion-de-lineas-y-estaciones-paradas-del-servicio-de-transportes-electricos.

[Des21] Secretaría de Desarrollo Urbano y Vivienda (SEDUVI). *Uso de Suelo*. Portal de Datos Abiertos CDMX, 2021. URL: https://datos.cdmx.gob.mx/dataset/uo-de-suelo.

A. Figuras omitidas

Población por alcaldía a través de los años

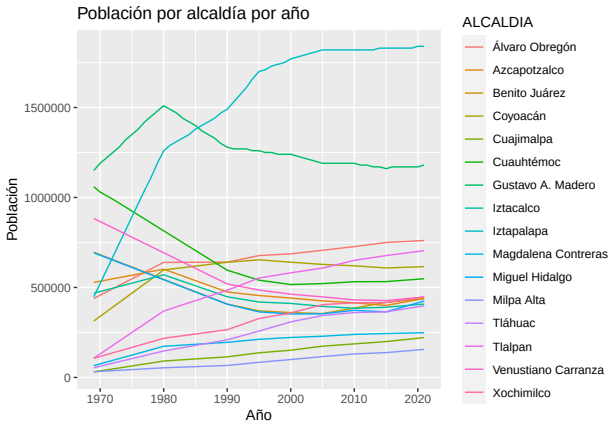


Figura 15: Población por alcaldía como función del tiempo

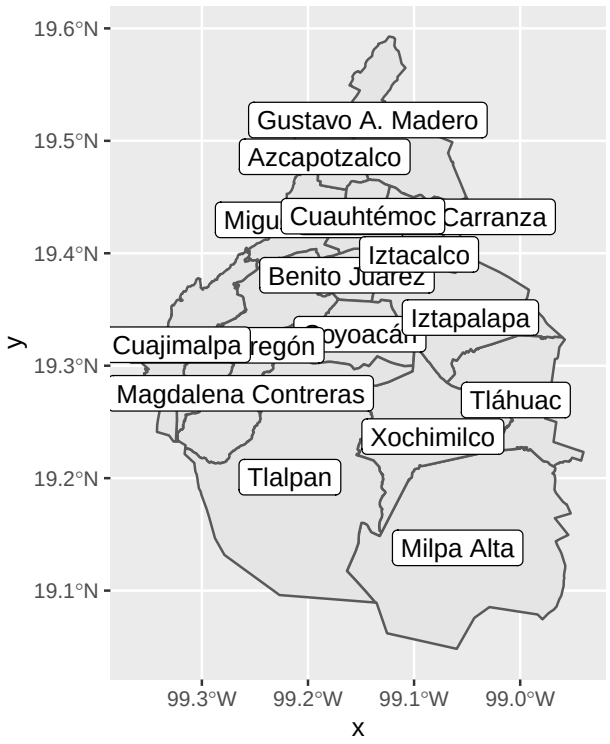


Figura 16: Mapa de la CDMX con división política.

B. Coeficientes de la regresión segmentada por alcaldía

A continuación la tabla a la que se hace referencia en la sección de regresión lineal. Contiene los coeficientes estimados para el modelo presentado para cada alcaldía junto con su error estándar, valor-p, y si es o no significativa a un nivel de 10 %.

Figura 17: Salida de la función `fa` aplicado al conjunto de datos

```
## Factor Analysis using method =  ml
## Call: fa(r = df2, nfactors = 1, rotate = "varimax", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##             ML1   h2   u2 com
## EST_TOTAL -0.68 0.47 0.53   1
## MEAN_DIST  0.68 0.47 0.53   1
##
##                 ML1
## SS loadings    0.93
## Proportion Var 0.47
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## The degrees of freedom for the null model are  1  and the objective function was  0.24 with
Chi Square of  3.06
## The degrees of freedom for the model are -1  and the objective function was  0
##
## The root mean square of the residuals (RMSR) is  0
## The df corrected root mean square of the residuals is  NA
##
## The harmonic number of observations is  15 with the empirical chi square  0  with prob <  NA
## The total number of observations was  15  with Likelihood Chi Square =  0  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  1.528
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 ML1
## Correlation of (regression) scores with factors   0.80
## Multiple R square of scores with factors          0.64
## Minimum correlation of possible factor scores     0.27
```

Tabla 6: Coeficientes estimados por alcaldía con significancia al 10 por ciento.

ALCALDIA	term	estimate	std.error	p.value	is.signif
Azcapotzalco	(Intercept)	4.840	14.848	0.746	no signif
Azcapotzalco	log(AÑO)	-0.335	1.872	0.859	no signif
Azcapotzalco	log(POBLACION)	-0.299	0.110	0.009	signif
Benito Juárez	(Intercept)	-1.180	5.614	0.834	no signif
Benito Juárez	log(AÑO)	0.329	0.722	0.651	no signif
Benito Juárez	log(POBLACION)	-0.132	0.027	0.000	signif
Coyoacán	(Intercept)	-3.028	3.211	0.350	no signif
Coyoacán	log(AÑO)	0.376	0.433	0.389	no signif
Coyoacán	log(POBLACION)	0.093	0.019	0.000	signif
Gustavo A. Madero	(Intercept)	-88.973	13.503	0.000	signif
Gustavo A. Madero	log(AÑO)	11.378	1.668	0.000	signif
Gustavo A. Madero	log(POBLACION)	0.426	0.166	0.013	signif
Iztacalco	(Intercept)	-9.473	5.969	0.119	no signif
Iztacalco	log(AÑO)	1.479	0.754	0.055	signif
Iztacalco	log(POBLACION)	-0.216	0.045	0.000	signif
Iztapalapa	(Intercept)	-95.934	15.174	0.000	signif
Iztapalapa	log(AÑO)	12.663	2.033	0.000	signif
Iztapalapa	log(POBLACION)	0.016	0.044	0.722	no signif
Magdalena Contreras	(Intercept)	3.139	2.234	0.166	no signif
Magdalena Contreras	log(AÑO)	-0.418	0.298	0.167	no signif
Magdalena Contreras	log(POBLACION)	0.048	0.007	0.000	signif
Miguel Hidalgo	(Intercept)	42.213	7.169	0.000	signif
Miguel Hidalgo	log(AÑO)	-5.294	0.922	0.000	signif
Miguel Hidalgo	log(POBLACION)	-0.246	0.033	0.000	signif
Milpa Alta	(Intercept)	-49.473	50.868	0.335	no signif
Milpa Alta	log(AÑO)	6.515	6.759	0.340	no signif
Milpa Alta	log(POBLACION)	0.049	0.112	0.663	no signif
Tláhuac	(Intercept)	-100.689	26.604	0.000	signif
Tláhuac	log(AÑO)	13.286	3.535	0.000	signif
Tláhuac	log(POBLACION)	0.002	0.050	0.962	no signif
Tlalpan	(Intercept)	-22.207	3.815	0.000	signif
Tlalpan	log(AÑO)	2.969	0.508	0.000	signif
Tlalpan	log(POBLACION)	-0.001	0.009	0.867	no signif
Venustiano Carranza	(Intercept)	109.372	17.626	0.000	signif
Venustiano Carranza	log(AÑO)	-13.903	2.262	0.000	signif
Venustiano Carranza	log(POBLACION)	-0.502	0.074	0.000	signif
Xochimilco	(Intercept)	-39.596	7.616	0.000	signif
Xochimilco	log(AÑO)	5.278	1.016	0.000	signif
Xochimilco	log(POBLACION)	-0.032	0.020	0.111	no signif