

Proyecto Estadística Aplicada III

Alonso Martinez Cisneros

Juan Carlos Sigler Priego
Esmeralda Altamirano

Carlos Delgado

2022-05-10

Índice

1 Propuesta de proyecto	1
2 Planteamiento del problema	1
3 Análisis exploratorio	2
3.1 Descripción de las variables de interés	2
3.2 Acceso a transporte con base en la población	5
4 PCA	13
5 Construcción de un índice de conectividad	14
5.1 Análisis del índice de conectividad por año	18
6 Regresión lineal	19
7 Interpretación, conclusiones, etc...	19

1 Propuesta de proyecto

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
- Metrobus
- Tren Ligero
- Cablebus

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 horas pico pasan convoyes nuevos de metro y metrobus 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 para 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] "Cuaajimalpa"       "Magdalena Contreras" "Milpa Alta"
## [16] "Álvaro Obregón"
```

Tenemos información para los siguientes años:

```
## [1] 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983
## [16] 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998
## [31] 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
## [46] 2014 2015 2016 2017 2018 2019 2020 2021
```

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

AÑO	ALCALDIA	POBLACION	MEAN_DIST	EST_TOTAL	ZOC_DIST
1969	Azcapotzalco	527857	2019.0774	0	2028.311
1970	Azcapotzalco	534554	1050.1630	0	2028.311
1971	Azcapotzalco	541251	1050.1630	0	2028.311
1972	Azcapotzalco	547948	958.8664	0	2028.311
1973	Azcapotzalco	554645	958.8664	0	2028.311
1974	Azcapotzalco	561342	958.8664	0	2028.311

3.1 Descripción de las variables de interés

Las variables son las siguientes:

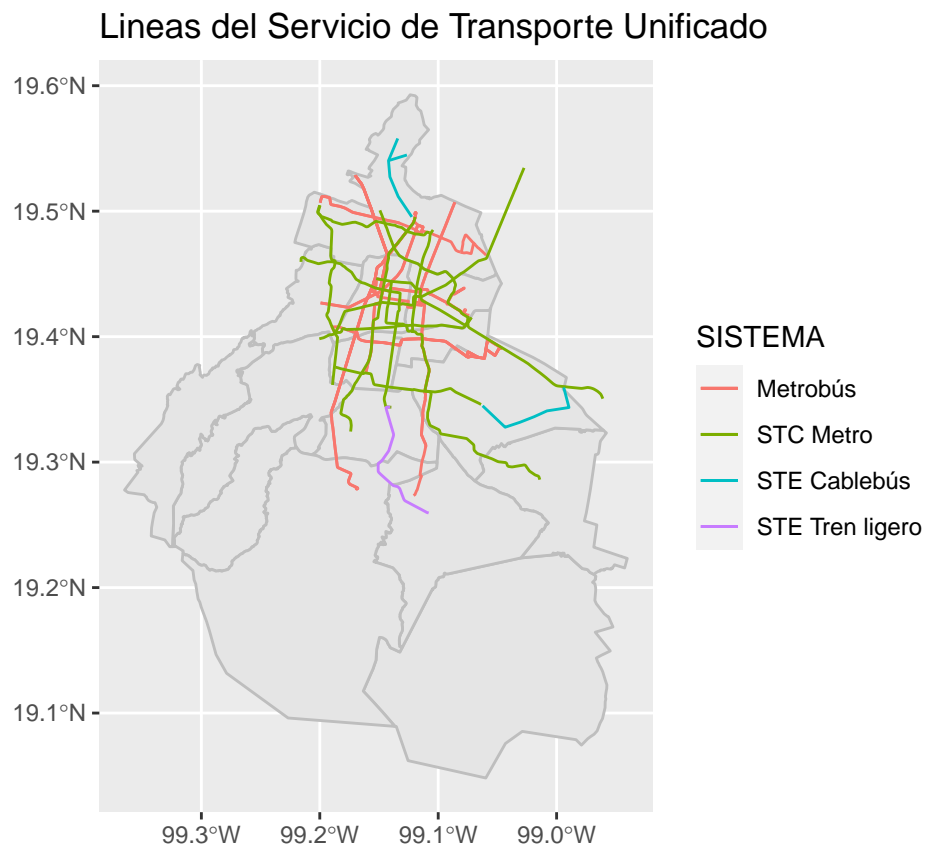
- AÑO: El año en el cual están medidas las variables.
- ALCALDIA: La demarcación territorial de delegación o Alcaldía al que corresponden los datos.

- **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.

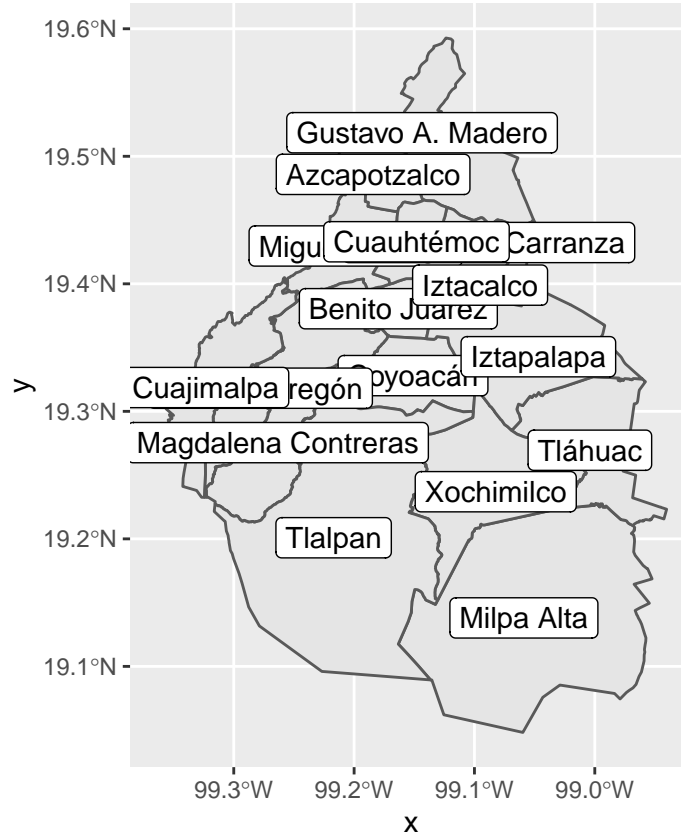
3.1.1 Número total de estaciones por delegación

Para encontrar el número de estaciones de transporte público masivo por delegación a un año dado utilizamos los conjuntos de datos [?, ?, ?]. En la figura ?? 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 sin no se conoce bien la ciudad. La mayor carencia aparente es al sur de la ciudad. En la figura ?? presentamos un mapa de la ciudad con divisiones políticas para hacer más fácil referirnos a alcaldías específicas.

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 [?]. 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 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 el valor real. ¿Por qué? Porque por un extremo tenemos la zona Observatorio y por la otra Desierto de los Leones. Desierto de los Leones 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

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.

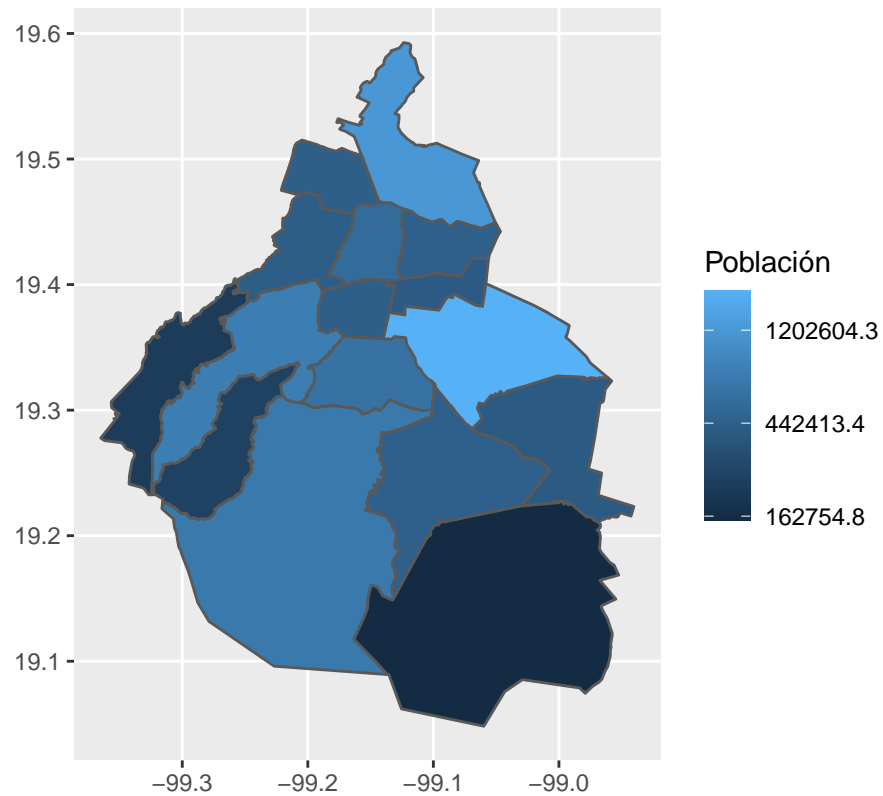
Reconocemos que la elección del zócalo de la ciudad como punto central puede parecer arbitraria. Sin embargo, nos parece justificable puesto que es una de las zonas más antiguas y por lo tanto el crecimiento de la zona metropolitana de la ciudad ha sido radialmente hacia afuera de esta zona. De manera 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 [] podemos ver cómo es la distribución geográfica de la población en la CDMX.

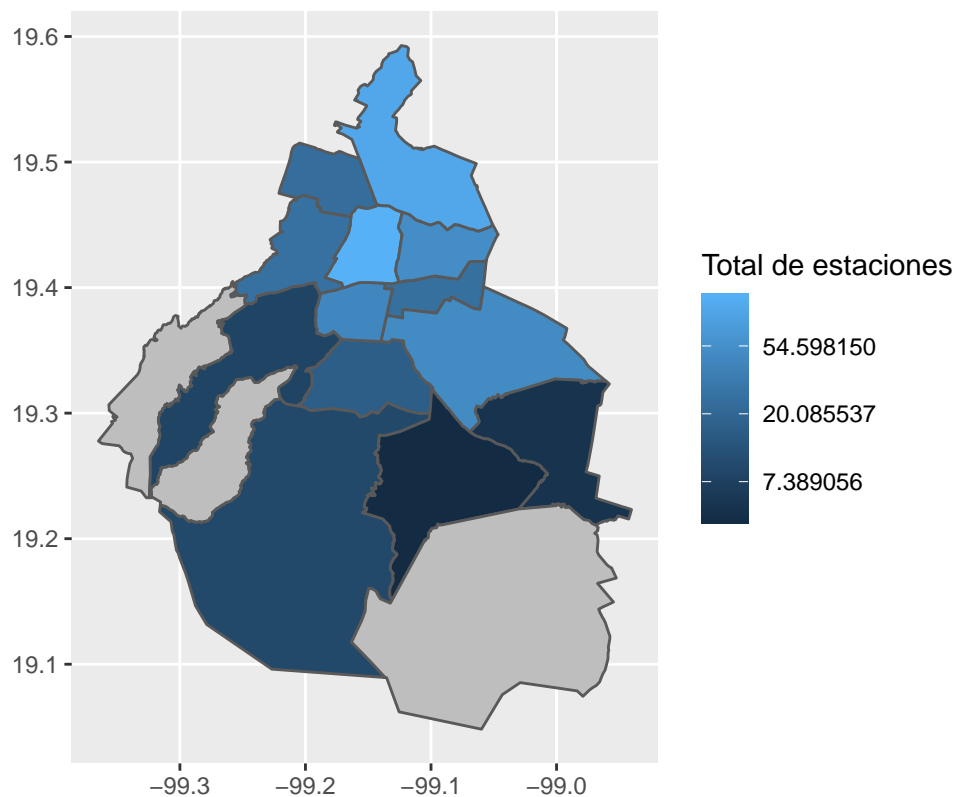
Alcaldías coloreadas por población al 2021



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 [] vemos las alcaldías coloreadas dependiendo de cuantas estaciones de transporte público al año 2021 tienen en total.

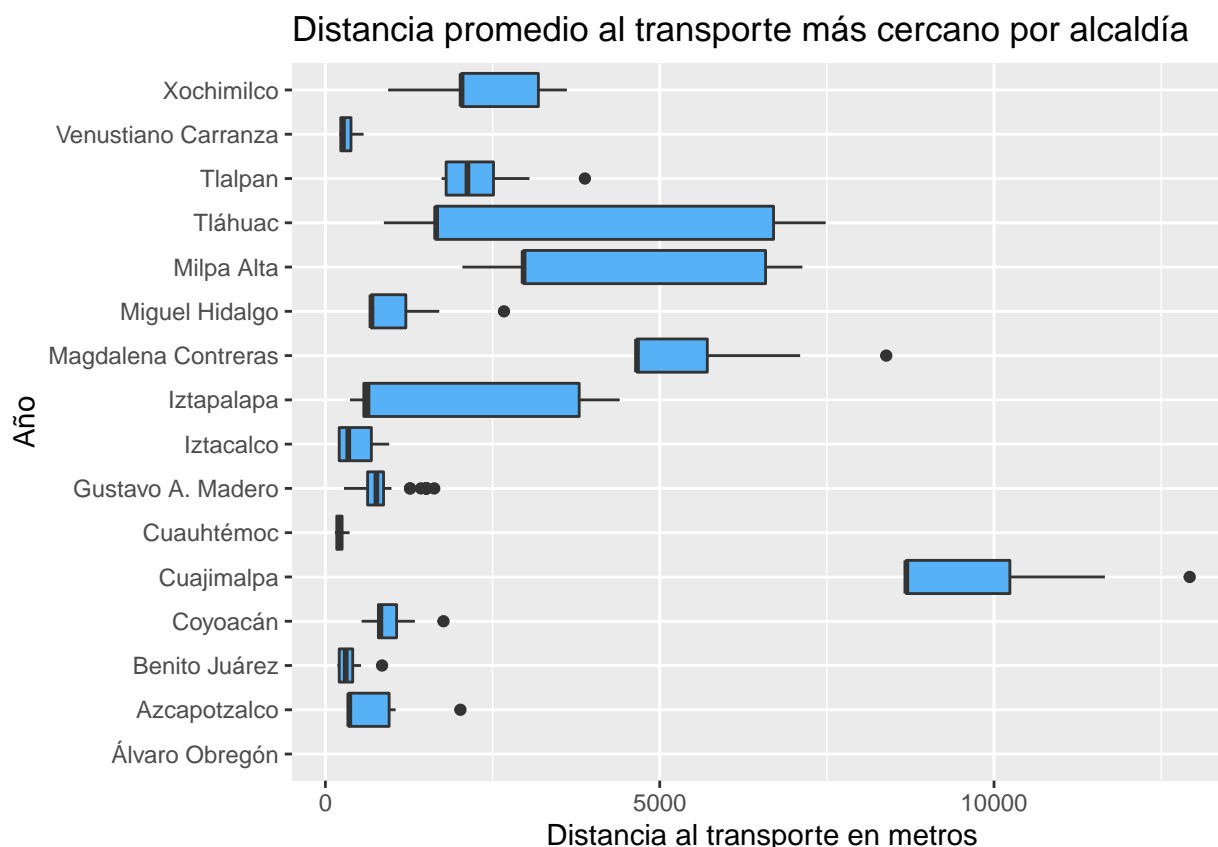
Total de estaciones por alcaldía.



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 [] 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.

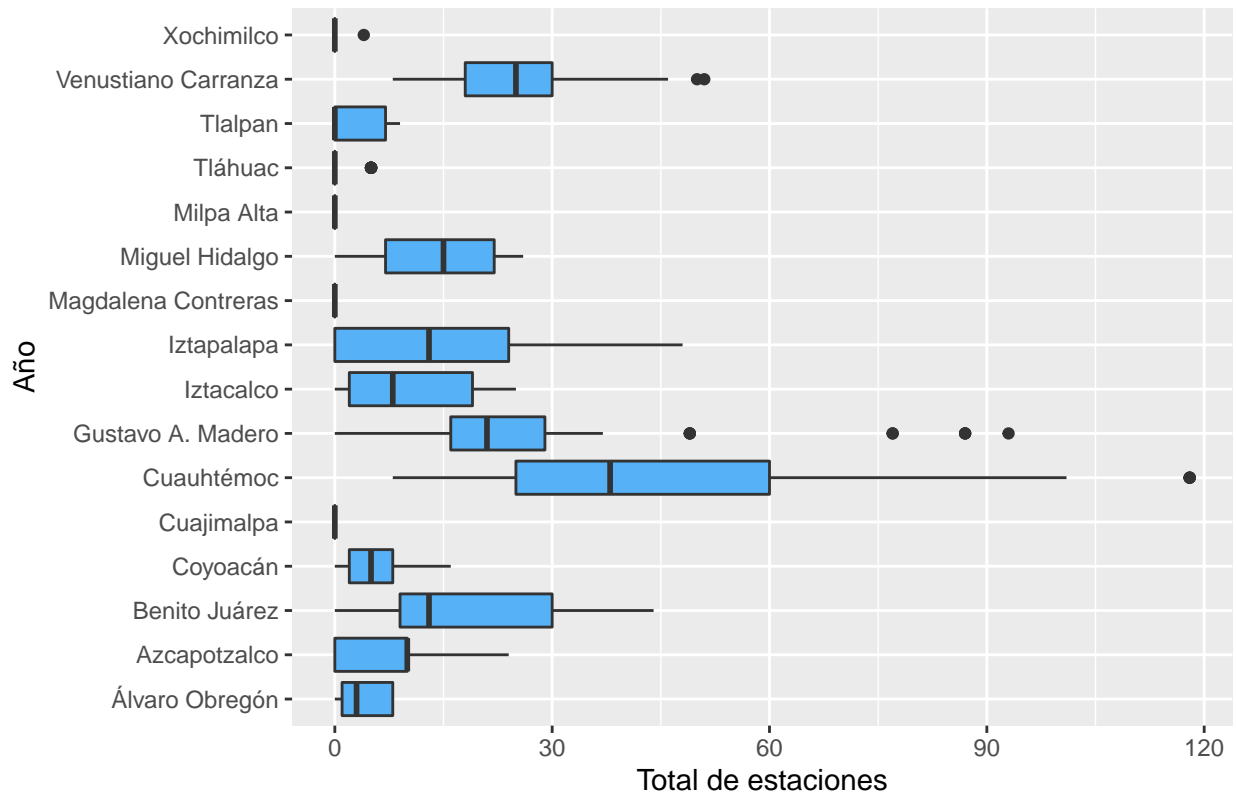


Podemos ver que las alcaldías de Cuauhtémoc, Benito Juárez y Venustiano Carranza tienen las menores varianzas en distancia media y además las más pequeñas. Estas alcaldías son precisamente la que definimos como “zona centro” desde el inicio. De esta observación confirmamos que la zona centro siempre 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 promedio a la estación de transporte más cercana sigue siendo muy baja en comparación a otras alcaldías, incluso las más pobladas como Iztapalapa.

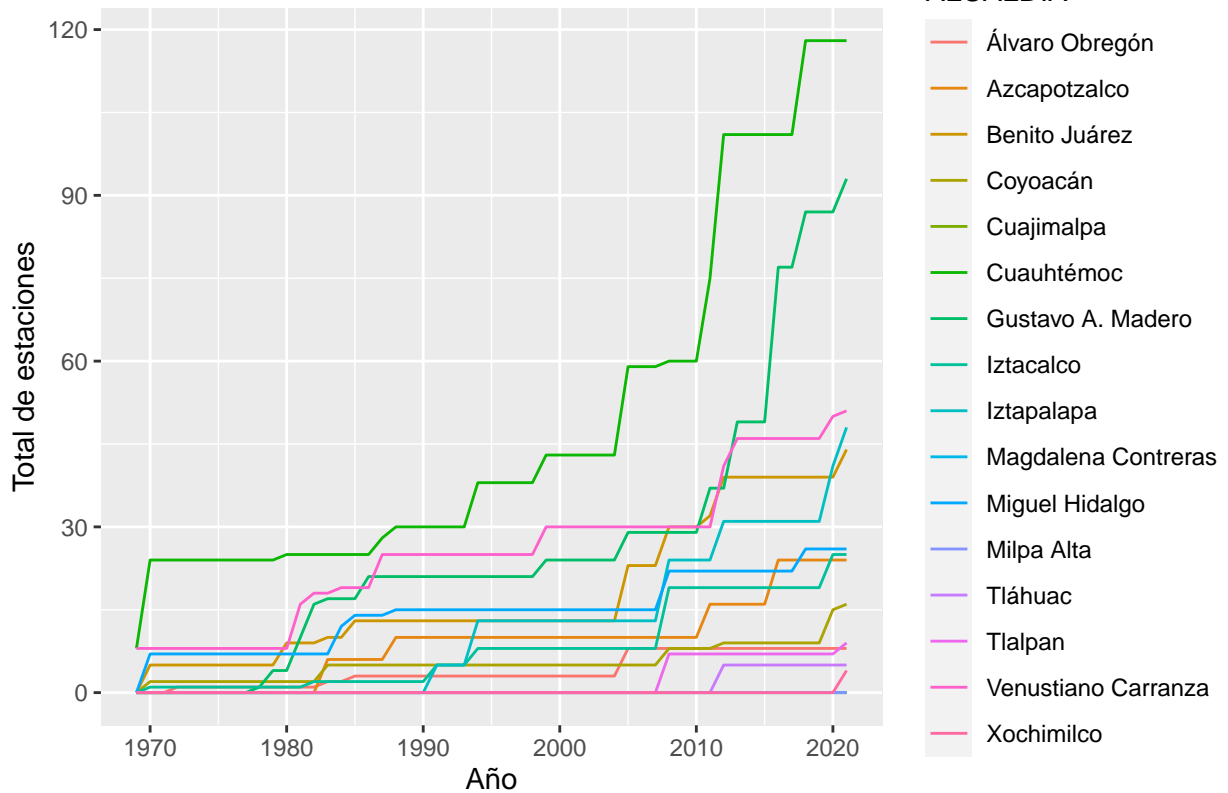
Por otro lado, Tláhuac, Cuajimalpa y Milpa Alta son los de mayor distancia y variación. En el caso de Tláhuac por ejemplo, siendo una de las alcaldías más al sur, lo que interpretamos es que su distancia promedio a las primeras estaciones era excesivamente alta y fue disminuyendo a medida que mejoró 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 muy rápidamente, lo cual sugiere que el sistema de transporte unificado evolucionó rápidamente para cubrir gran parte de la zona metropolitana.

Total de estaciones por alcaldía a través del tiempo.



Total de estaciones por alcaldía por año.

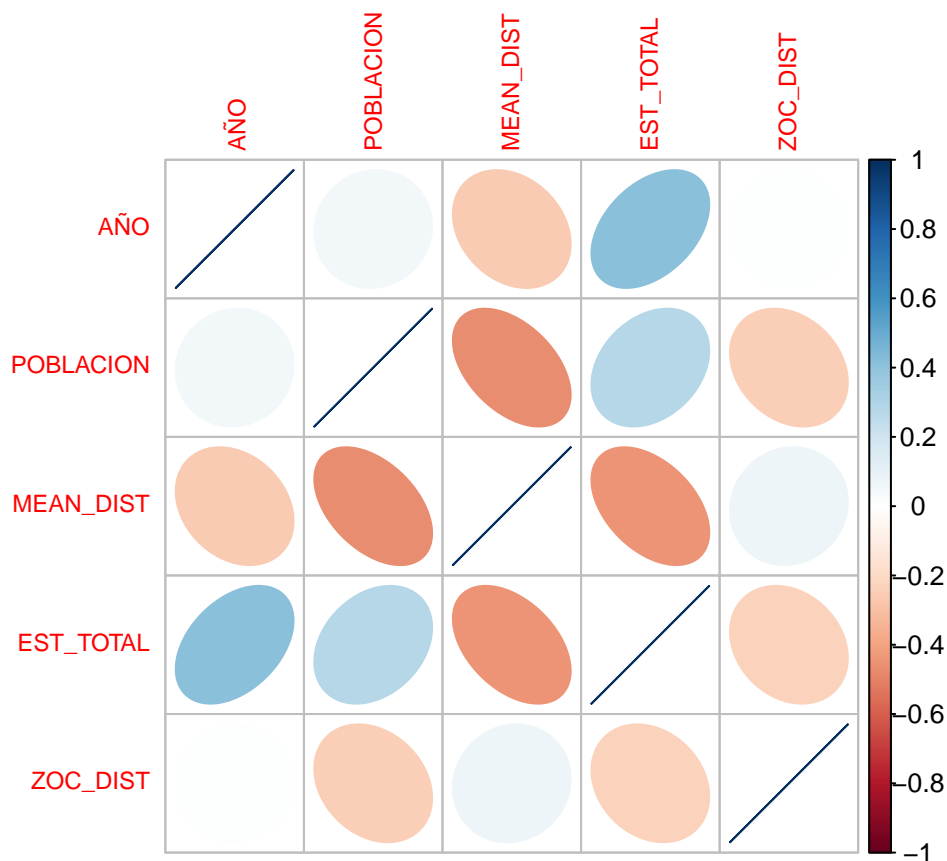


En la figura [] consideramos el número total de estaciones en la alcaldía como función del tiempo. Aquí podemos ver que el número de estaciones en las alcaldías de la zona centro excede vastamente el de las alcaldías más periféricas, como Iztapalapa. Analizando la variabilidad mediante el ancho de la 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 han hecho han estado en la zona circundante y han mejorado su conectividad. Esa zona es precisamente la zona centro. Una pista más que indica que la inversión en creación de nuevas líneas ha estado privilegiando a la zona centro.

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. En la figura [] vemos un diagrama de correlación para las variables estudiadas.



Efectivamente se cumple que la correlación de distancia al Zocalo 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

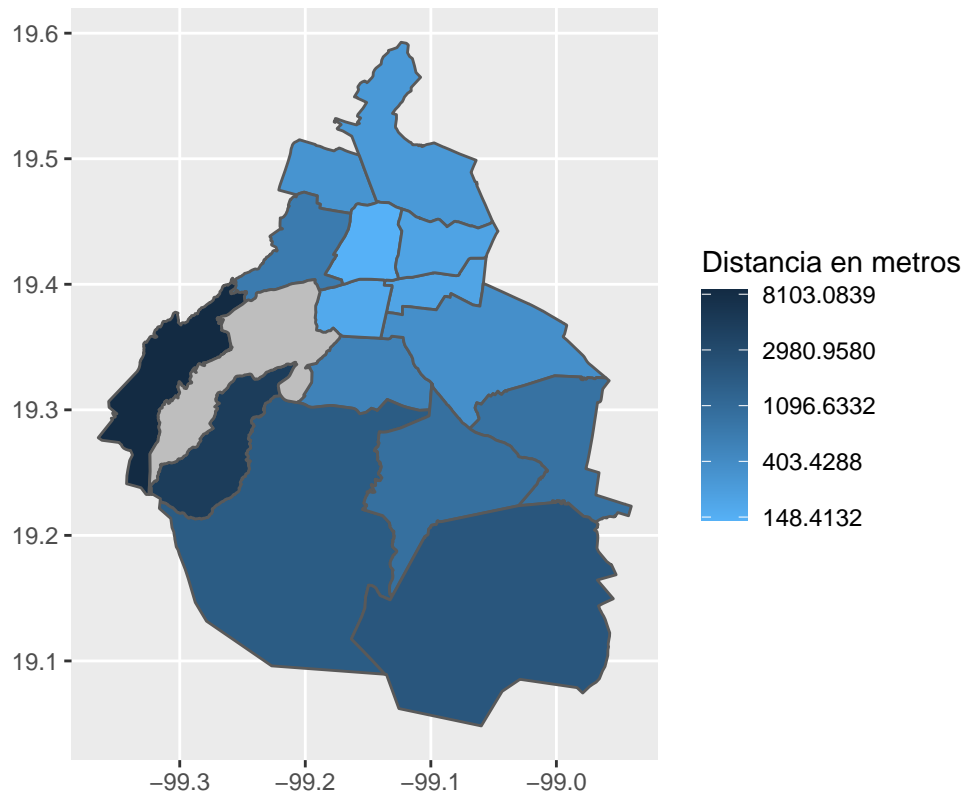
entre distancia al zócalo con el número total de estaciones. Es decir, entre más lejos está la alcaldía del zócalo menor es el número total de estaciones a las que se tiene acceso. Las correlaciones son aparentemente débiles, pero notables. Sospechamos que la correlación se hace más fuerte a medida que se va hacia atrás en el tiempo cuando había menos estaciones en total. El corolario es que esta conectividad si ha estado mejorando desde que se empezó a construir la primera línea de metro hasta la actualidad.

La matriz explícita de correlaciones es:

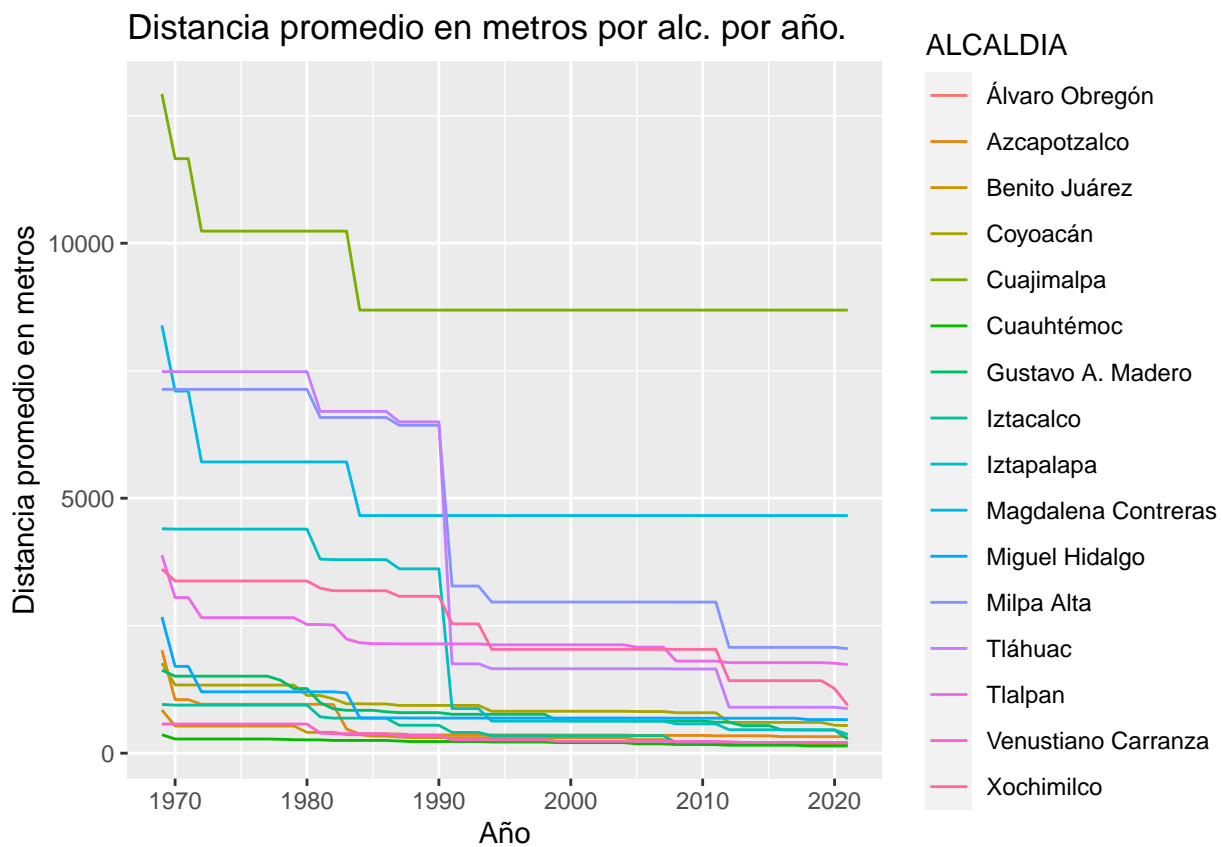
	AÑO	POBLACION	MEAN_DIST	EST_TOTAL	ZOC_DIST
AÑO	1.0000000	0.0502763	-0.2532690	0.4107378	0.0000000
POBLACION	0.0502763	1.0000000	-0.4607899	0.2897473	-0.2489191
MEAN_DIST	-0.2532690	-0.4607899	1.0000000	-0.4426642	0.0766026
EST_TOTAL	0.4107378	0.2897473	-0.4426642	1.0000000	-0.2217675
ZOC_DIST	0.0000000	-0.2489191	0.0766026	-0.2217675	1.0000000

Si graficamos la distancia promedio al sistema de transporte por alcaldía la correlación espacial entre distancia al centro de la ciudad aproximado mediante la posición del zócalo podremos tener indicación visual de si nuestra hipótesis tiene sentido. 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 el 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 con el corredor Insurgentes Sur respectivamente. Por otro lado, las zonas como Los Dínamos y Las Águilas están muy lejos del resto del sistema.

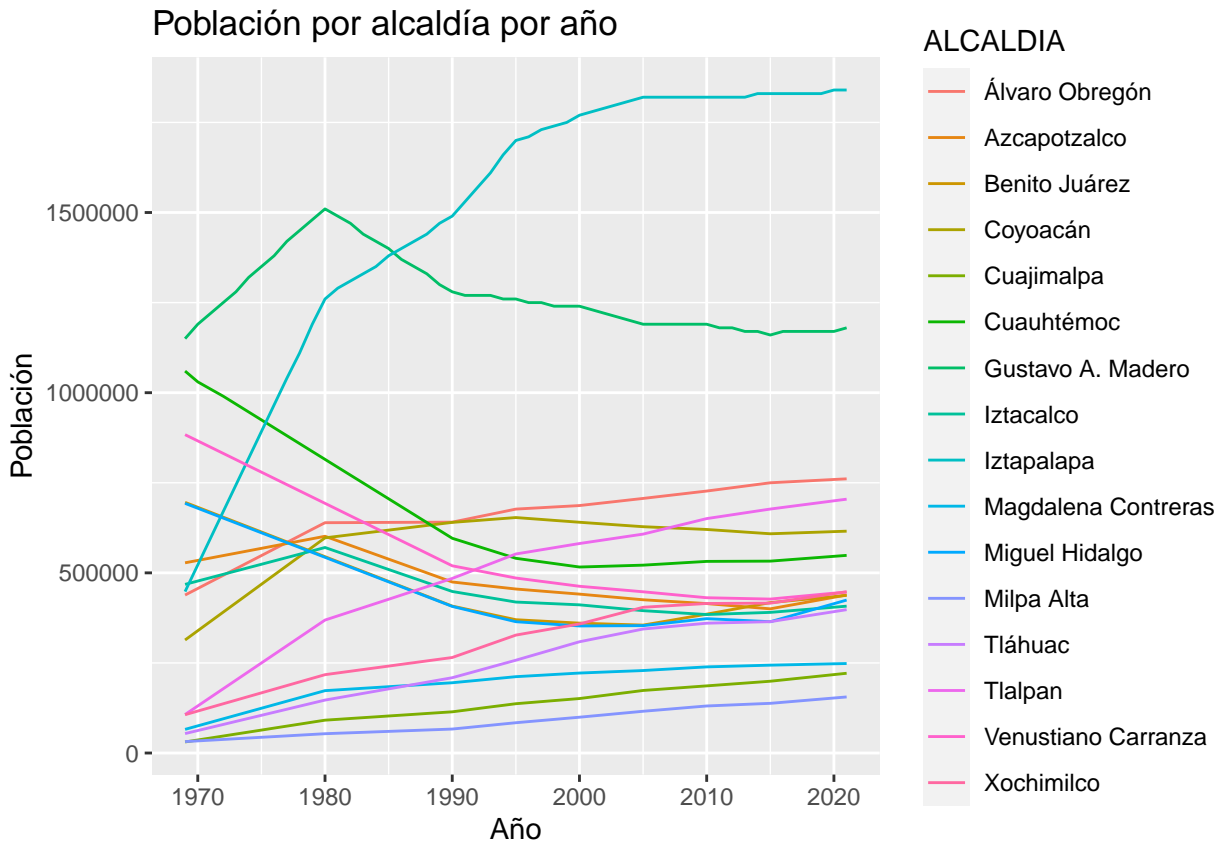
Distancia promedio al transporte más cercano



Ahora vemos cómo evoluciona como función del tiempo.

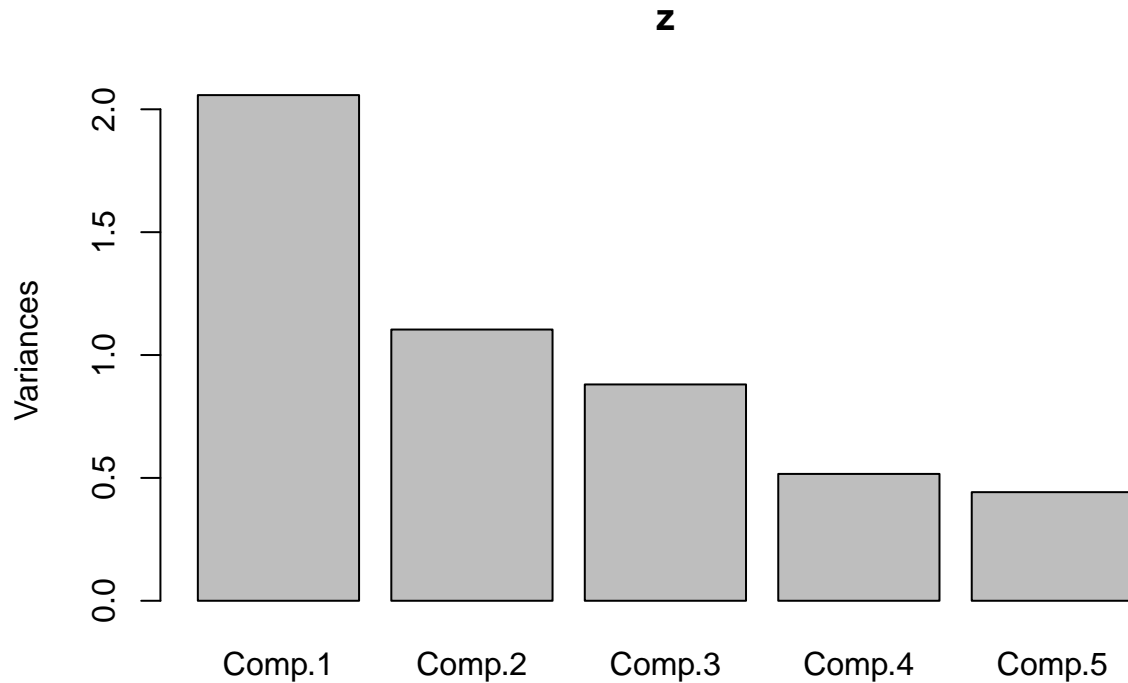


[Aqui grafica perrona de correlación a través del tiempo.]



4 PCA

```
## Importance of components:
##               Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
## Standard deviation  1.434476  1.050592  0.9382218  0.7185053  0.66483939
## Proportion of Variance 0.411544  0.2207517  0.1760520  0.1032500  0.08840228
## Cumulative Proportion 0.411544  0.6322957  0.8083478  0.9115977  1.00000000
##
## Loadings:
##               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
```



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 y la prueba Kaiser–Meyer–Olkin (KMO) indica una adecuación medianamente regular.

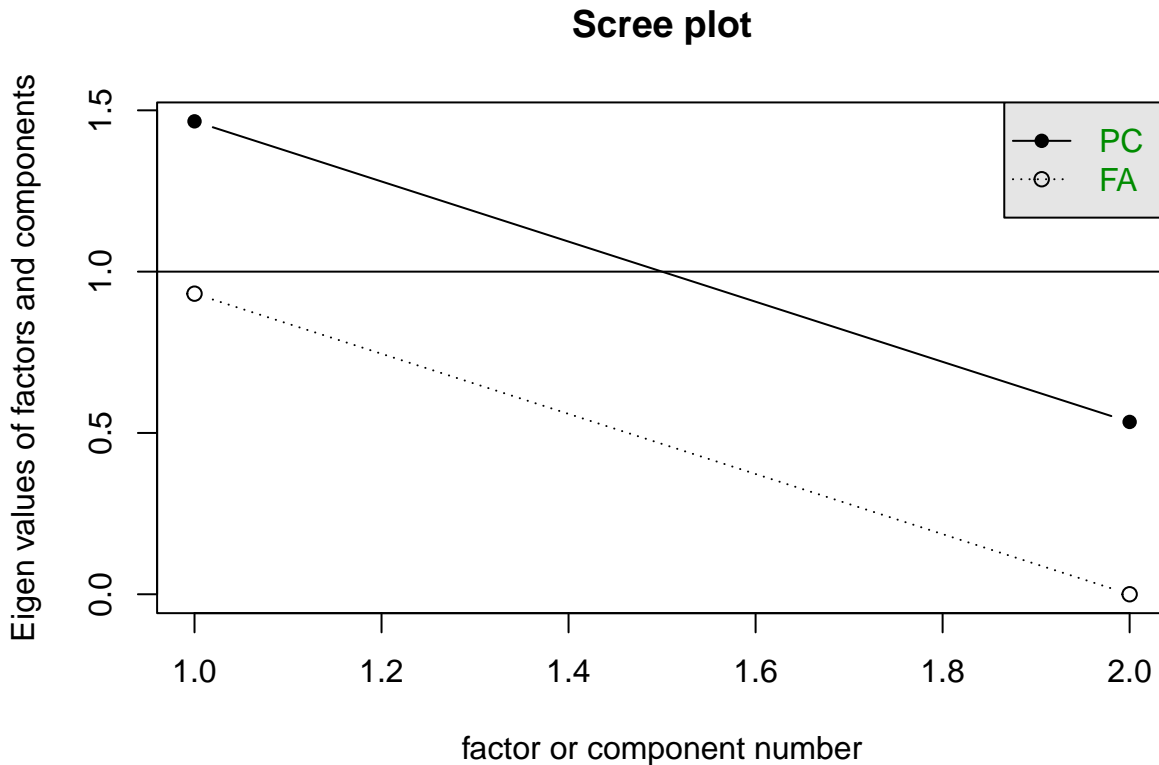
Las pruebas de Bartlett dice lo siguiente:

```
##
## Bartlett test of homogeneity of variances
##
## data: df2
## Bartlett's K-squared = 94.484, df = 1, p-value < 2.2e-16
```

Por otro lado, KMO dice lo siguiente:

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor(df2))
## Overall MSA = 0.5
## MSA for each item =
## EST_TOTAL MEAN_DIST
## 0.5 0.5
```

El gráfico de sedimentación (scree plot en inglés) indica que un factor es suficiente en este caso.

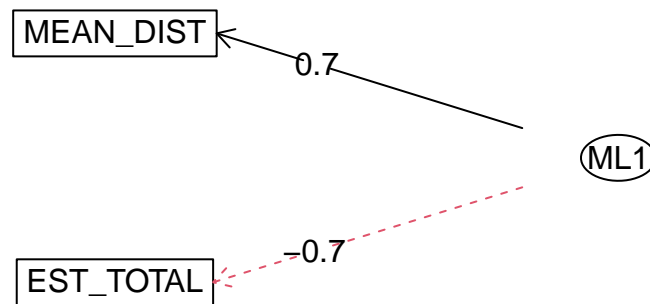


Al ver el modelo generado, vemos que el factor o constructo contrasta la población y cantidad de estaciones contra la distancia media a las mismas. Un valor muy alto del índice indicaría que tienes mucha población o estaciones, mientras que un índice bajo indicaría una valor mucho mayor de la distancia respecto a las estaciones y a la población.

```
## 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 = 0.24
## 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
##
## Correlation of (regression) scores with factors    ML1    0.80
## Multiple R square of scores with factors          0.64
## Minimum correlation of possible factor scores      0.27
```

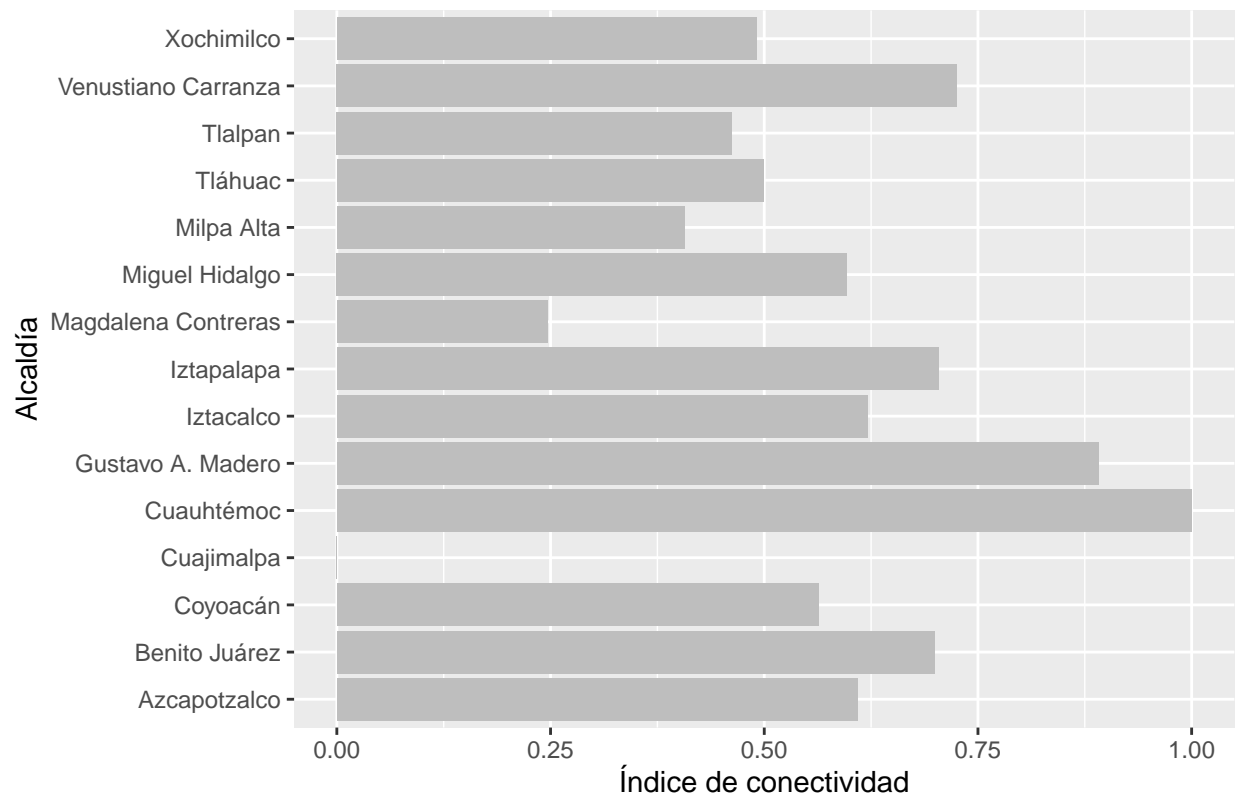
Factor Analysis



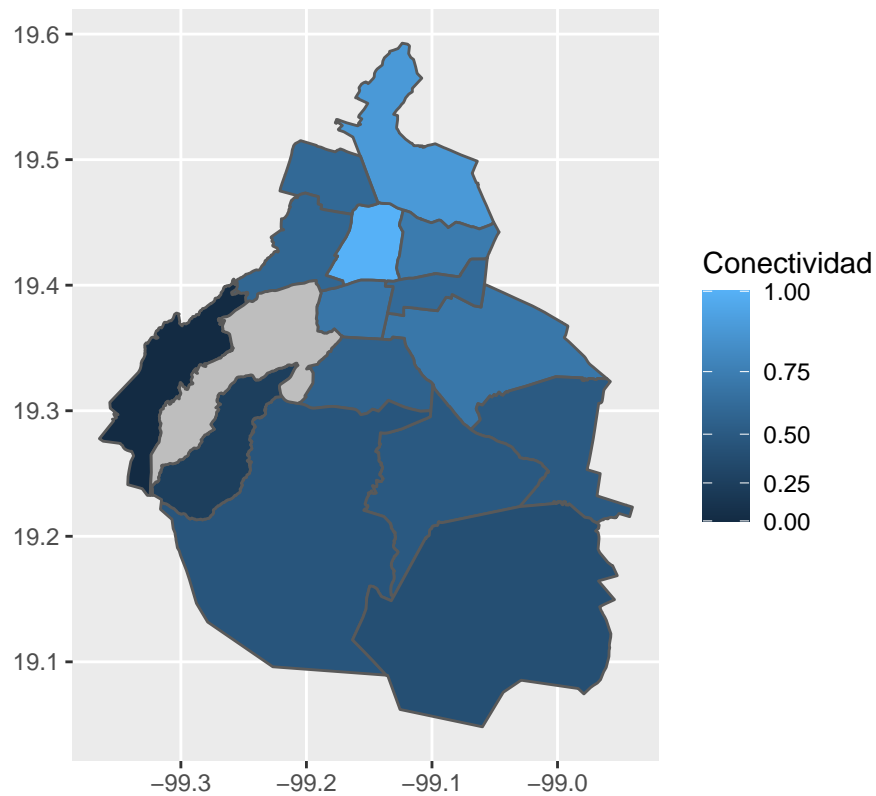
Construimos ahora el índice para cada delegación en 2021. Podemos ver que las delegaciones con el índice más alto coinciden con las delegaciones con más estaciones y más población, que son las de la zona centro e Iztapalapa.

	ML1
Cuauhtémoc	1.0000000
Gustavo A. Madero	0.8907132
Venustiano Carranza	0.7253886
Iztapalapa	0.7035597
Benito Juárez	0.6990018
Iztacalco	0.6207869
Azcapotzalco	0.6093896
Miguel Hidalgo	0.5969240
Coyoacán	0.5637512
Tláhuac	0.4989385
Xochimilco	0.4910410
Tlalpan	0.4621720
Milpa Alta	0.4067971
Magdalena Contreras	0.2468354
Cuajimalpa	0.0000000

Índice de conectividad por alcaldía al 2021

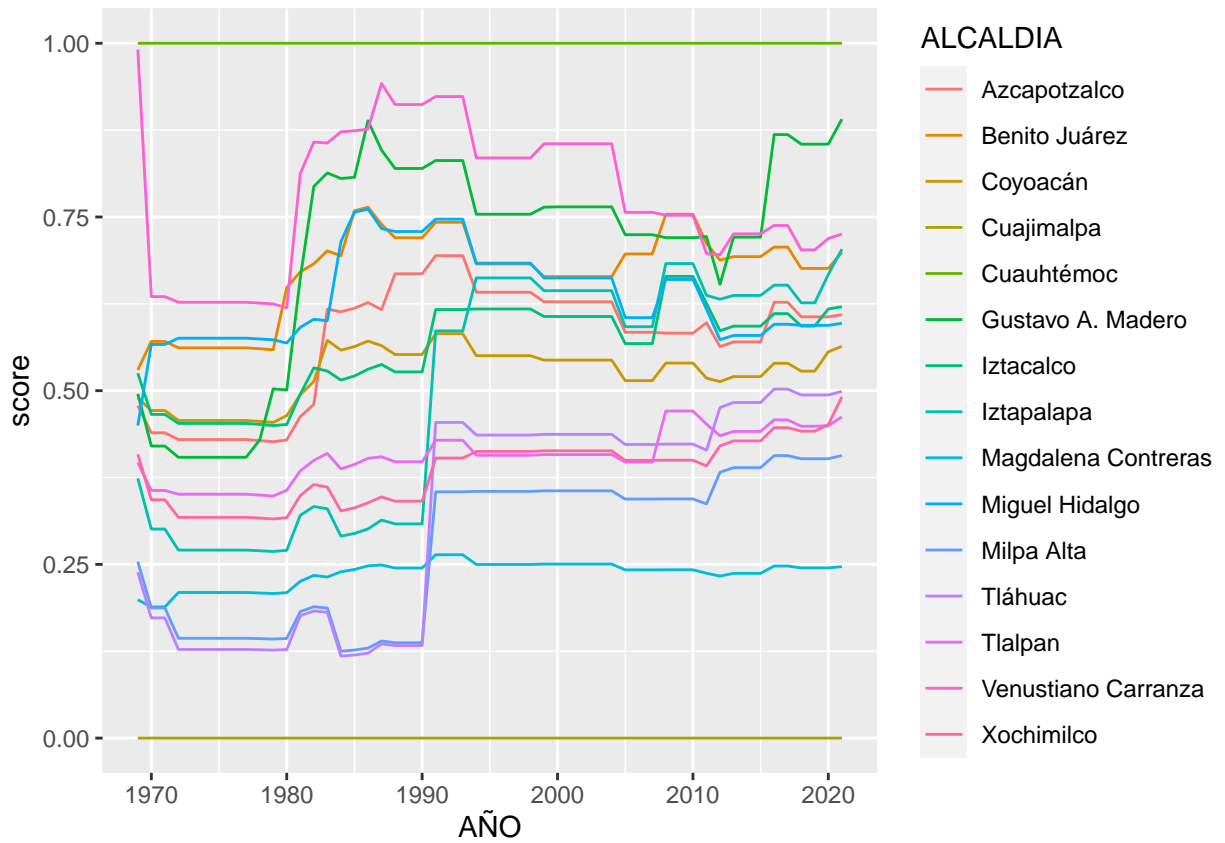


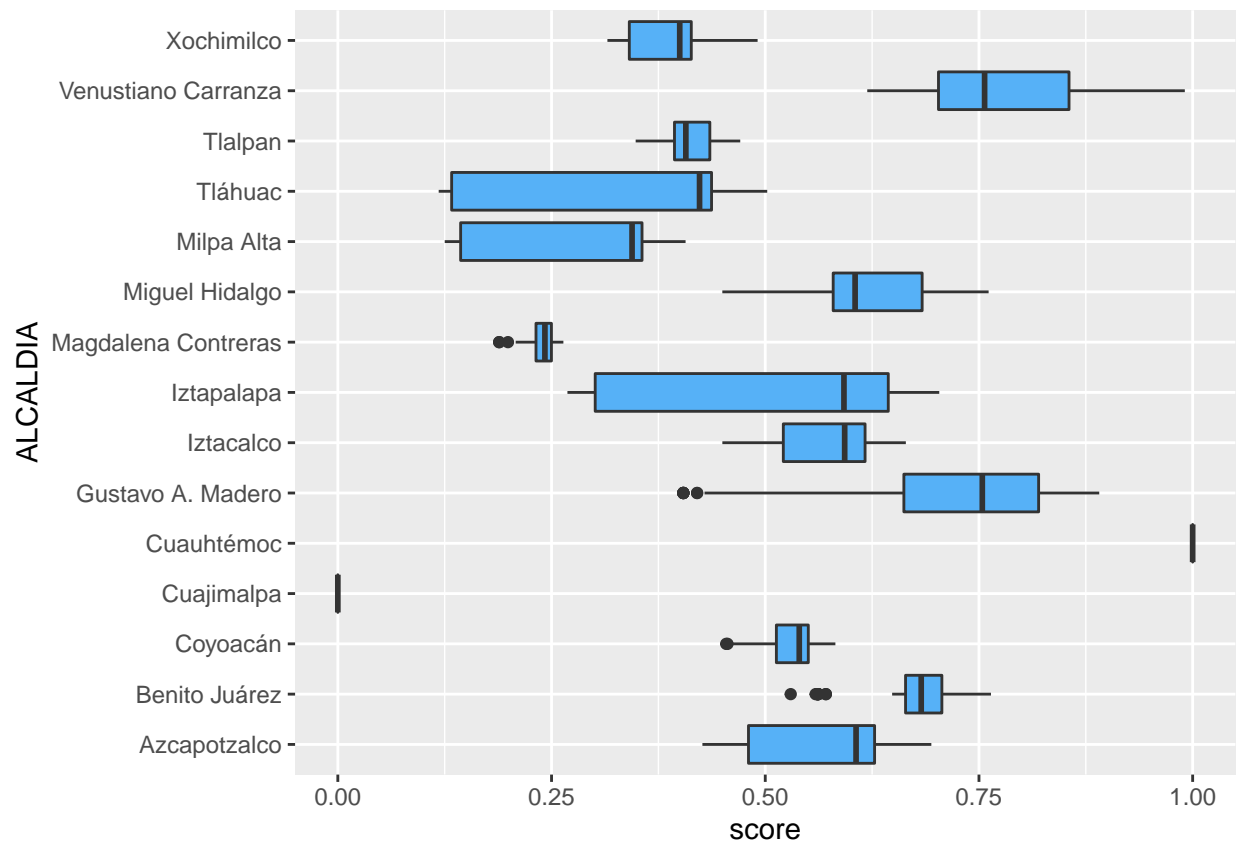
Alcaldías por conectividad al 2021



5.1 Análisis del índice de conectividad por año

Queremos ver cómo ha evolucionado el índice de conectividad de las diversas alcaldías a medida que ha ido creciendo el sistema de transporte unificado.





6 Regresión lineal

7 Interpretación, conclusiones, etc...