Universidad del Rosario

Facultad de Economía

Escuela Doctoral de Economía

Maestría en Economía

Tesis:

La educación secundaria y sus dos dimensiones, efectos del barrio y del colegio sobre los resultados saber 11.

Jacobo Rozo Alzate

Asesores de tesis:

María José Álvarez

Darío Maldonado

ÍNDICE

Resumen:

Este trabajo estudia los resultados en matemáticas y lenguaje de 32000 estudiantes en la prueba saber 11 del 2008, de la ciudad de Bogotá. Este análisis reconoce que los individuos se encuentran contenidos en barrios y colegios, pero no todos los individuos del mismo barrio asisten a la misma escuela y viceversa. Con el fin de modelar esta estructura de datos se utilizan varios modelos econométricos, incluyendo una regresión jerárquica multinivel de efectos cruzados. Nuestro objetivo central es identificar en qué medida y que condiciones del barrio y del colegio se correlacionan con los resultados educacionales de la población objetivo y cuáles características de los barrios y de los colegios están más asociadas al resultado en las pruebas. Usamos datos de la prueba saber 11, del censo de colegios c600, del censo poblacional del 2005 y de la policía metropolitana de Bogotá. Nuestras estimaciones muestran que tanto el barrio como el colegio están correlacionados con los resultados en las pruebas; pero el efecto del colegio parece ser mucho más fuerte que el del barrio. Las características del colegio que están más asociadas con el resultado en las pruebas son la educación de los profesores, la jornada, el valor de la pensión, y el contexto socioeconómico del colegio. Las características de los barrios más asociadas con el resultado en las pruebas son, la presencia de universitarios en la UPZ, un clúster de altos niveles de educación y nivel de crimen en el barrio que se correlaciona negativamente. Los resultados anteriores fueron hallados teniendo en cuenta controles familiares y personales.

Introducción:

La educación siempre ha sido una parte vital de las sociedades humanas, es la forma por la cual se trasmite toda la acumulación de conocimiento que se ha adquirido por las generaciones anteriores; conocimiento que es en ultimas la herramienta más poderosa del ser humano. Pero en la sociedad moderna cobra aún más importancia, convirtiéndose en la forma más importante de ascensión social, es por esto que los estudios sobre educación han cobrado tanta importancia. De igual manera encontrar los factores que afectan el aprendizaje de un niño o un joven no es nada sencillo, estos dependen en gran medida de cualidades personales y familiares, el estudio de Patacchini y Zenou (2009), muestra que no solo importa el contexto del estudiante sino también la disposición y actitud de los padres hacia el aprendizaje de su hijo. Dado lo anterior, es importante usar herramientas adecuadas para poder analizar los diferentes contextos a los que se ve expuesto el individuo, y así encontrar indicios sobre las desigualdades educativas, es por esto que además de controlar por condiciones familiares es válido tener en cuenta cuál es el efecto del colegio en los resultados educativos y la posibilidad de acción a ese nivel; y por otro lado tener en cuenta el efecto del vecindario del individuo y analizar qué condiciones del vecindario favorecen o perjudican el rendimiento educativo.

Siguiendo nuestra idea, podemos decir que los estudios sobre educación no pueden solo tener en cuenta los factores que repercuten directamente en la educación, como la educación de los profesores o la educación de los padres, también deben tener en cuenta variables contextuales como la violencia del barrio (J. Hardign, 2009). Pues ya sea por efecto de los pares, por presión colectiva del grupo social o por un modelo de vida, es muy posible que condiciones del vecindario también afecten los resultados educacionales de los individuos. En esta línea, encontramos el trabajo de Garner y Radenbush (1991) quienes encuentran que tras controlar por la habilidad y condiciones educativas del hogar y de la escuela de un individuo, las condiciones del vecindario, como mayor desempleo o presencia de pobreza, afectan negativamente los resultados educativos.

A priori hay muchos factores en juego a la hora de predecir los resultados en educación de un estudiante. Nuestra hipótesis es que los factores contextuales son muy importantes, en particular nos enfocamos en las condiciones contextuales del colegio y del barrio las cuales posiblemente están asociadas con el aprendizaje en edades escolares. En términos más precisos, nuestro ejercicio consiste en estudiar la contribución del barrio y del colegio sobre los resultados en matemáticas y lenguaje de una prueba estandarizada que presentan todos los estudiantes colombianos al finalizar la secundaria (prueba Saber 11 del ICFES). El ejercicio impone un reto metodológico: poder modelar los resultados educativos teniendo en cuenta una doble anidación (en barrios y colegios) cruzada, lo cual se refiere a que no todos los estudiantes de un mismo barrio asisten al mismo colegio y viceversa.

Más específicamente, no existe un consenso sobre qué condiciones de los colegios promueven o detienen el progreso educativo de un estudiante; por ejemplo no se sabe si una composición etaria o niveles socioeconómicos más homogéneos son positivos o negativos para los resultados educativos (Hoxby y Weingarth, 2006; Owens, 2010). Sobre las condiciones del vecindario, el debate está aún más abierto, no se sabe si la segregación es positiva o negativa, o incluso si el barrio tiene algún efecto sobre la educación en lo absoluto (Hoxby, Olmo y Weingarth, 2013; Goux y Maurin, 2003; Kaztman y Retamoso, 2007).

Nuestra estructura de datos exige que tengamos en cuenta simultáneamente la anidación de los datos en barrios, colegios y los cruces entre barrios y colegios, es por esto que elegimos los modelos jerárquicos multi nivel de efectos cruzados; entrando en detalle, la estructura educativa en Bogotá nos permite medir el efecto vecindario en estudiantes que habitan y no habitan en el vecindario de su institución educativa, esto es un avance respecto a otros estudios de educación que tengan en cuenta las condiciones del barrio, pues en la mayoría de estos los estudiantes habitan en el mismo barrio donde se ubica el colegio. No tener en cuenta esta estructura podría llevarnos a calcular erróneamente la varianza del modelo (Raudenbush y Bryck, 2002). Además nos permite suponer que ciertas variables, como la educación de los padres, no tienen el mismo efecto para los distintos grupos, hallando una pendiente distinta para cada uno de ellos. Nuestro método de estimación, modela los tres grupos, colegio, barrio

y los cruces entre colegio y barrio, con un término aleatorio similar al error, ahondaremos en este aspecto más adelante.

El problema es computacionalmente muy exigente, lo que hace necesario una aproximación parsimoniosa. Por esto proponemos una estimación en dos etapas. En la primera etapa, descomponemos la varianza del resultado en la prueba estandarizada en el componente atribuible al barrio y al colegio, controlando por condiciones familiares. En la segunda etapa, tomamos estas varianzas y buscamos las variables de cada una de las dos agrupaciones (barrios o colegios) que están más correlacionadas con la misma varianza. En nuestra base de datos tenemos variables familiares tomadas del cuestionario adjunto a la prueba saber11, también tenemos variables a nivel de barrio tomadas del censo del 2005 y de la policía metropolitana, y para describir los colegios tenemos variables del censo a colegios c600.

Las preguntas centrales de esta investigación son: 1) si existe o no un efecto del vecindario en los resultados educativos de los jóvenes; 2) Identificar cuál de las dos anidaciones tenidas en cuenta dada nuestra estructura de datos, barrios y colegios, explica una mayor porción de la varianza; y 3) que variables tanto de los colegios como de los barrios afectan los resultados en educación.

Nuestros principales resultados van en la misma línea de la literatura, sobre las variables familiares podemos ver que la educación de los padres y el nivel de ingresos afectan positivamente los resultados en las pruebas, además estudiantes en extra edad y estudiantes con hermanos que abandonaron el colegio tienen peores resultados. Por otra parte, el nivel educativo de los profesores y la educación agregada de los padres impactan positivamente los resultados del colegio.

Encontramos que si existe un efecto del vecindario en la educación, pero el efecto del colegio es mucho mayor, entrando en detalle, para lenguaje el vecindario explica casi un8.9% de la varianza mientras el colegio explica aproximadamente un 30%, para la prueba de matemáticas encontramos que el barrio explica un 4.1% mientras el colegio explica un 34% de la varianza de la variable dependiente, lo anterior sumado a otros resultados nos lleva a pensar que la prueba de lenguaje es más susceptible a las condiciones sociales del individuo; de igual manera el nivel individual y familiar son los más importantes, en lenguaje explican un 57% de la varianza de esta prueba y en matemáticas explican un 56%. Hallamos evidencia a favor de un efecto espacial del vecindario, en otras palabras que la agrupación de vecindarios con mayores niveles de educación (clúster) genera un efecto del barrio positivo. También podemos ver que una mayor proporción de universitarios en la UPZ genera mejores condición para la educación a nivel de barrio, mientras que fenómenos como el crimen son perjudiciales.

1. Estadísticas descriptivas

1.1 Descripción de la base de datos:

Nuestra base de datos se compone de 4 encuestas, la encuesta principal es el cuestionario socioeconómico adjunto a la prueba saber 11; a esta le agregamos información promedio del censo de colegios c600 que usamos para describir las instituciones educativas. A esta base le extraemos los datos atípicos de la muestra, en otras palabras se eliminaron de la muestra los estudiantes con un resultado de cero en las pruebas de lenguaje o matemáticas, los estudiantes mayores de 26 años, los estudiantes quienes no registraban ningún dato en la educación de alguno de los padres y los estudiantes que asisten a jornadas nocturnas o sabatinas. Por último, para describir el vecindario de los individuos usamos el censo poblacional del 2005 y los datos de la policía metropolitana de Bogotá. La siguiente tabla nos muestra el número de observaciones según se van incluyendo las bases:

Tabla 1:

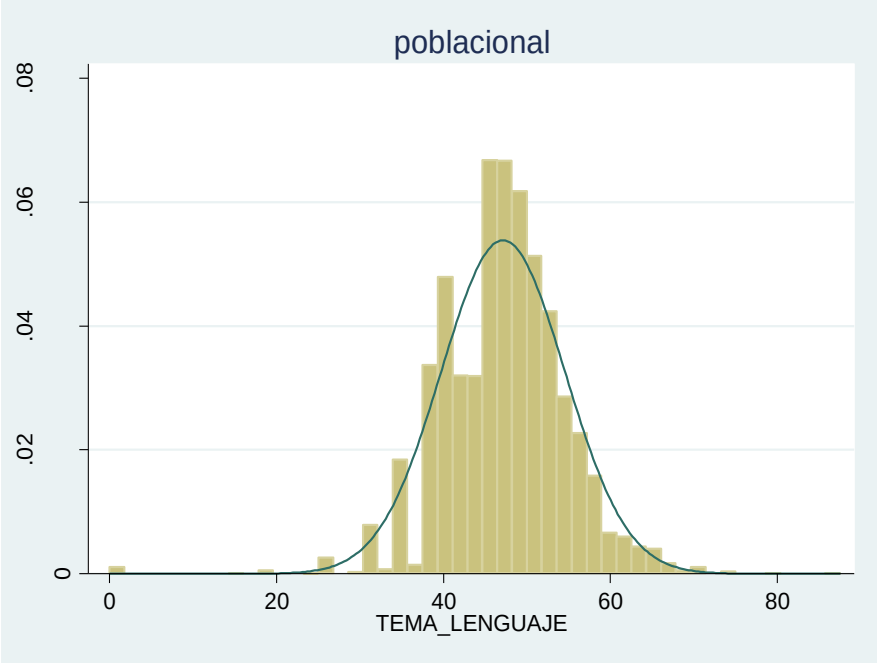
Construcción de las bases de datos.

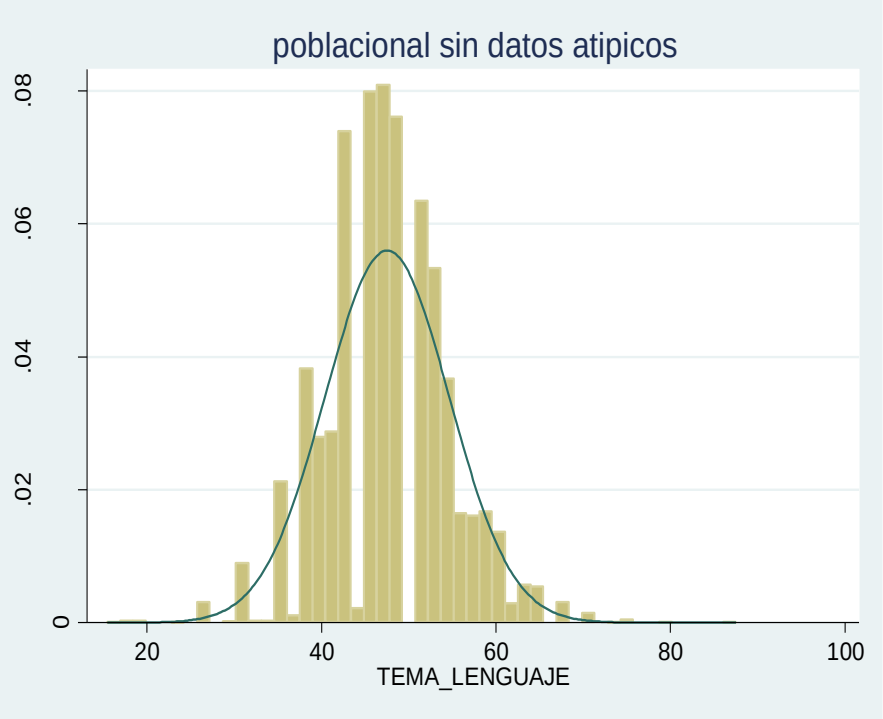
Bases de datos	Numero de observaciones
sb11	95766
c600	91232
sin datos atípicos	76536
geo-referenciados (censo)	32692

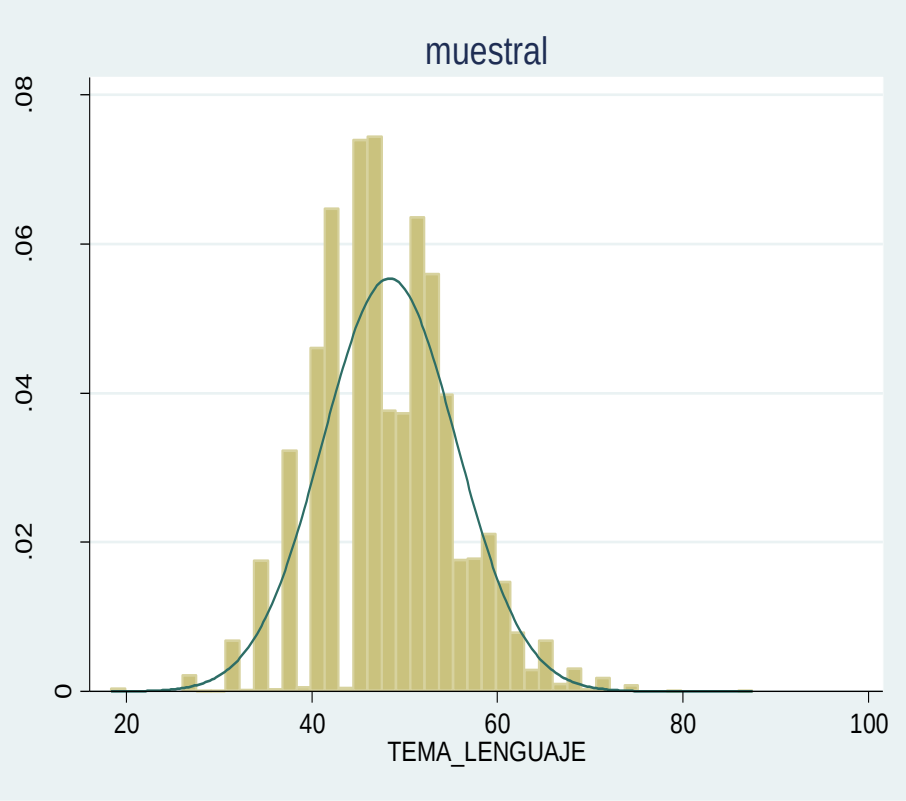
Como mostramos anteriormente perdemos una buena cantidad de datos por la presencia de datos atípicos y por el proceso de geo-referenciar, por esto es posible que las variables de interés tengan una distribución distinta, para esto compararemos los histogramas de las variables dependientes de la muestra poblacional (sin incluir los datos atípicos y los otros individuos excluidos de la muestra) con nuestra muestra:

Grafica 1: histogramas de las variables dependientes.

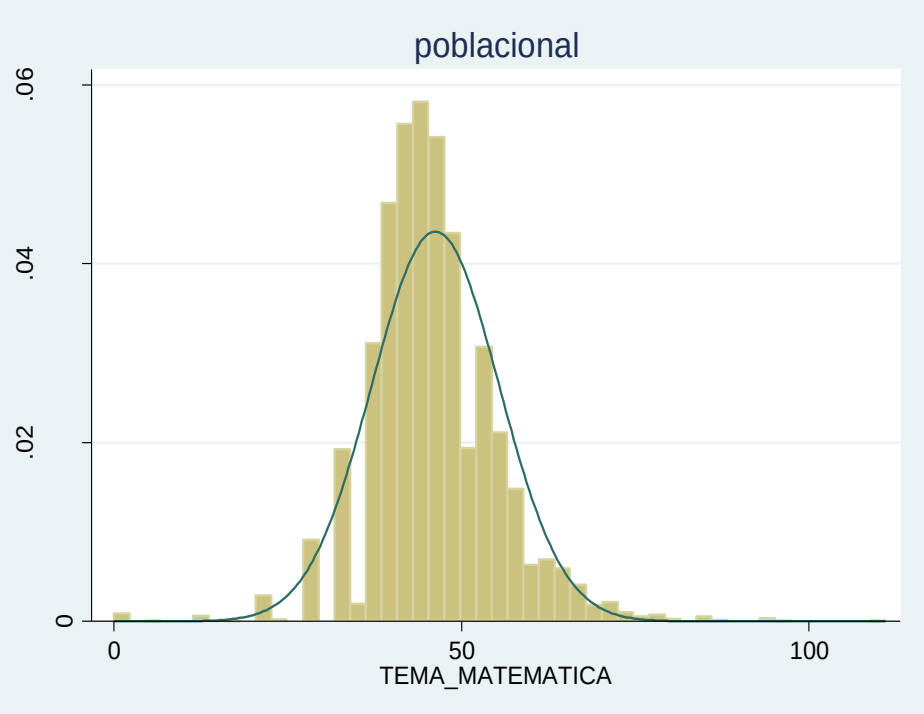
Lenguaje:

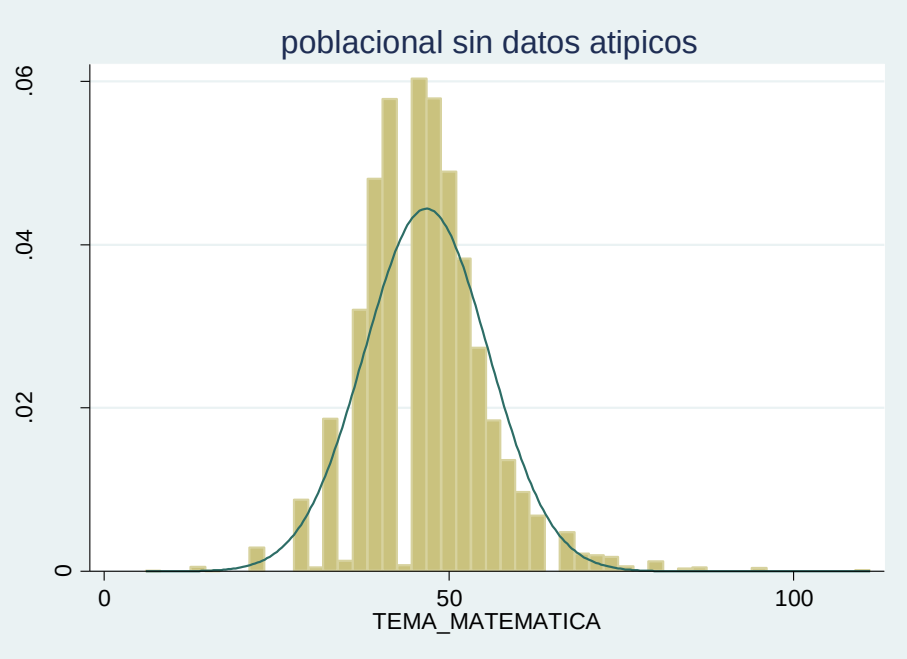

poblacional
.08
.06
.04
.02
0
0
20
40
60
80
TEMA_LENGUAJE

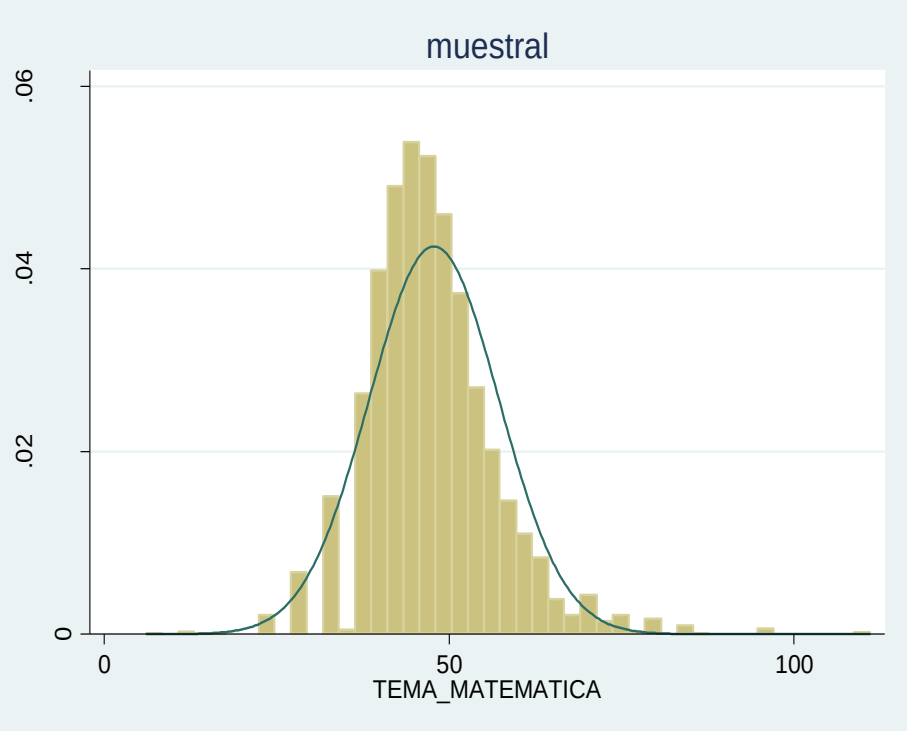

poblacional sin datos atipicos
.08
.06
.04
.02
0
20
40
60
80
100
TEMA_LENGUAJE



Matemáticas:







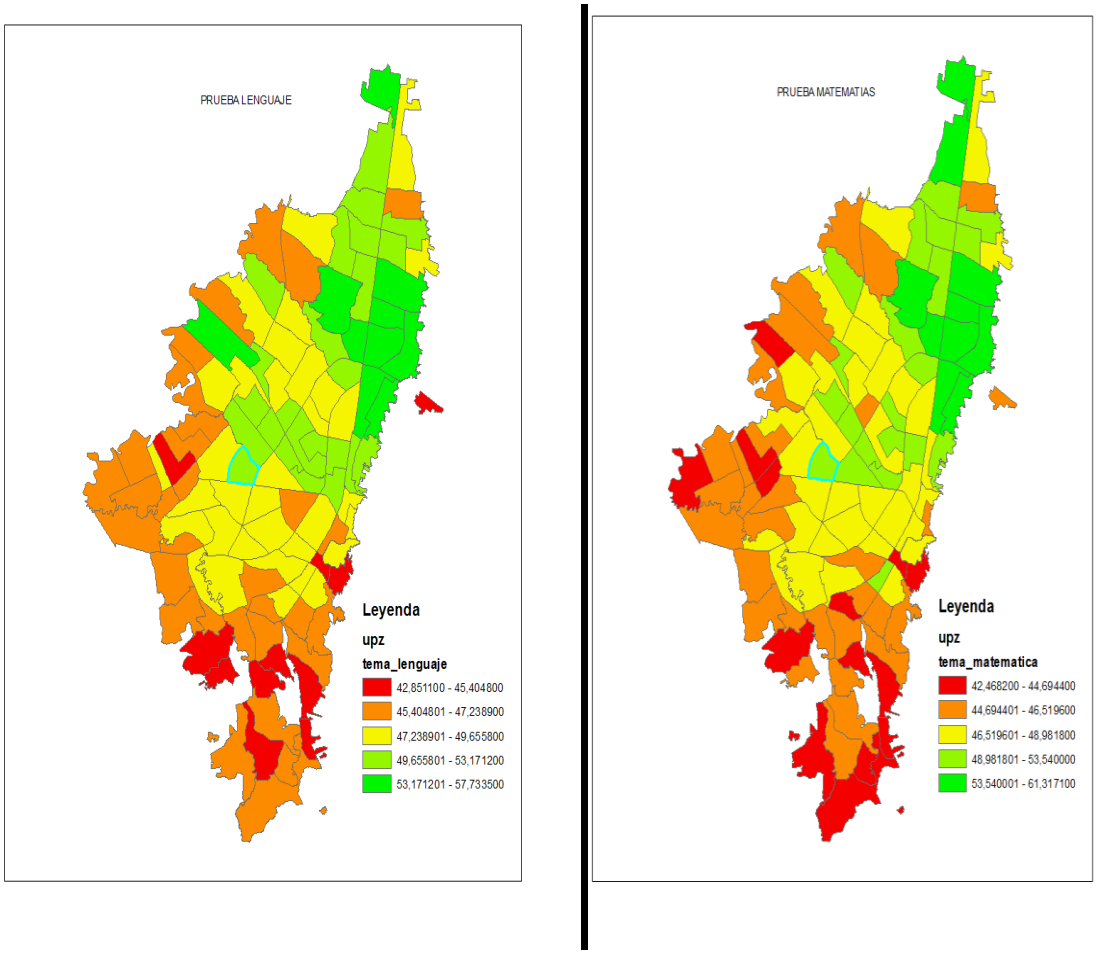
Como podemos ver en los histogramas la diferencia entre la base poblacional y nuestra sub-muestra no es radical, en especial en la prueba de lenguaje, dado esto podemos suponer que nuestra perdida de datos no implica mayor sesgo. De igual manera, realizamos varios ejercicios de regresión comparando la base poblacional, la base poblacional sin datos atípicos

y la base muestral. Los resultados de esta comparación no muestran mayor diferencia, la magnitud y la significancia de los estimadores es similar; incluso la varianza explicada por la anidación de los datos en colegios es prácticamente igual, para ver una descripción más detallada de esto referirse al anexo 1.

1.2 Análisis gráfico:

A continuación analizamos la distribución espacial entre los resultados en las pruebas de lenguaje y matemáticas, para esto utilizaremos mapas de la distribución de estas pruebas en Bogotá:

Grafica 2: Resultados saber11 en Bogotá.



En los mapas anteriores podemos ver un efecto de segregación norte sur respecto al resultado en ambas pruebas, este resultado es similar al de Aliaga y Álvarez (2010), quienes encuentran que efectivamente existe segregación residencial a niveles grandes de agregación en Bogotá, no solo en términos educativos, sino también en términos laborales, donde el norte concentra

población mejor educada y con mejores condiciones socioeconómicas. Dada esta estructura de vivienda en la ciudad, es válido pensar que se deben tener en cuenta zonas que agrupan mejores y peores condiciones, y que no solo la pertenencia a estos barrios influye, sino también la pertenencia a esta aglomeración de barrios puede tener efectos sobre el logro educativo.

1.3 Variables:

Las variables incluidas en nuestro modelo se dividen en tres grupos, primero están las variables a nivel familiar e individual que son tomadas de la encuesta socioeconómica adjunta a la prueba Saber 11. En este grupo de variables tenemos la educación de los padres medida en años de educación, que según la literatura se muestra como uno de los principales predictores de la educación del individuo (Kaztman y Retamoso, 2007; Raudenbush, 1993). También tenemos una dummy que indica si el estudiante tiene hermanos y si tienen hermanos menores de 15 años que no están estudiando, la presencia de hermanos en el hogar puede correlacionarse con mejores resultados educacionales (Bordalejo y Calero, 2009) y el nivel educativo de los mismos es una buena proxy del ambiente educativo en el hogar. Entre estas variables también tenemos variables dicotómicas que indican la ocupación de los jefes de hogar y el ingreso mensual del hogar, un mayor ingreso y ocupaciones de alto estatus suelen tener una relación positiva en el logro educativo (Gabiria y Barrientos, 2001; Garner y Raudenbush, 1991). A continuación mostramos una descripción de las variables del hogar (ver anexo 2 para la comparación con la base poblacional):

Tabla 2:

Descripción de las variables.

Variable	Mean	Std. Dev.
Matemática	47.80095	9.391547
Lenguaje	48.36536	7.203706
Sexo	0.4698091	0.4990953
edad centrada	-0.0336284	1.122711

hermano desertor	0.0404992	0.1971299
no hermanos	0.3389208	0.4733501
educación centrada padre	0.1282121	4.398419
educación centrada de la madre	0.0711038	4.061902
ingreso familiar centrado	0.0543573	1.539583
madre con otra ocupación	0.0521211	0.2222744
padres con otra ocupación	0.0582711	0.234259
madres estudiantes	0.0033953	0.0581713
padres estudiantes	0.0008259	0.0287269
madres que viven de la renta	0.0043436	0.0657635
padres que viven de la renta	0.0047106	0.0684732
madres jubiladas	0.0161457	0.1260378
padres jubilados	0.04271	0.2022053
madres que se ocupan en el hogar	0.3757267	0.4843169
padres que se ocupan en el hogar	0.0069361	0.0829948
padres cuenta propia	0.2105854	0.4077303

madres cuenta propia	0.1145743	0.3185121
madres independientes profesionales	0.040868	0.1979872
padres independientes profesionales	0.1028895	0.3038189
madres auxiliares	0.0697059	0.2546543
padres auxiliares	0.0556899	0.2293251
madres empleadas técnicas o profesionales	0.0744258	0.2624664
padres empleadas técnicas o profesionales	0.0794912	0.2705077
madres directoras	0.0393139	0.1943435
padres directores	0.0563806	0.2306585
madres gerentes	0.0209808	0.1433221
padres gerentes	0.0367812	0.1882269
madres pequeñas empresarias	0.0602084	0.2378759
padres pequeños empresarios	0.0835492	0.2767145
madres empresarias	0.0213838	0.144662
padres empresarios	0.0412709	0.198919

Variable	Mean	Std. Dev.	Min
matematica	47,70142	9,357431	6,08
lenguaje	48,29043	7,188256	18,33
sexo	0,4704714	0,4991345	0
edan centrada	-0,01885	1,140865	-4,144879
m_priminc	0,0831463	0,2761071	0
h_priminc	0,1088471	0,3114517	0
m_primaria	0,1168192	0,3212093	0
h_primaria	0,1328786	0,3394484	0
m_secundaria	0,2530075	0,4347415	0
h_secundaria	0,2349335	0,4239634	0
m_tecinc	0,0276291	0,16391	0
h_tecinc	0,0186496	0,1352862	0
m_tecnico	0,0793185	0,2702391	0
h_tecnico	0,0605537	0,2385133	0
m_profinc	0,0401773	0,1963777	0
h_profinc	0,034709	0,1830445	0
m_profecional	0,1558453	0,3627138	0
h_profecional	0,1691706	0,3749079	0
m_postgr	0,0491855	0,2162583	0
h_postgr	0,0589133	0,2354657	0
hermano_desertor	0,0418178	0,2001754	0
no_hermano	0,3363265	0,4724589	0
cent_ing_fam	0,0105104	1,531444	-2,103143
m_otraocup	0,0521211	0,2222744	0
h_otraocup	0,0549704	0,2279257	0
m_estudian	0,0032522	0,0569358	0
h_estudian	0,0007771	0,0278655	0
m_renta	0,0042595	0,0651265	0
h_renta	0,0044897	0,0668558	0
m_jubilados	0,0161457	0,1260378	0
h_jubilados	0,04271	0,2022053	0
m_hogar	0,3757267	0,4843169	0
h_hogar	0,0069361	0,0829948	0
h_cuentapropia	0,2105854	0,4077303	0

m_cuentapropia	0,1145743	0,3185121	0
m_indep_prof	0,040868	0,1979872	0
h_indep_prof	0,1028895	0,3038189	0
m_auxiliar	0,0697059	0,2546543	0
h_auxiiar	0,0556899	0,2293251	0
m_emple_tecprof	0,0744258	0,2624664	0
h_emple_tecprof	0,0794912	0,2705077	0
m_direc	0,0393139	0,1943435	0
h_direc	0,0563806	0,2306585	0
m_gernetes	0,0209808	0,1433221	0
h_gerentes	0,0367812	0,1882269	0
m_peqempre	0,0602084	0,2378759	0
h_peqempre	0,0835492	0,2767145	0
m_empresarios	0,0213838	0,144662	0
h_empresarios	0,0412709	0,198919	0
educacion centrada padre	0,6446561	2,821774	-11,67242
educacion centrada padre	0,401237	2,533227	-12,03772

Teniendo en cuenta nuestro método, para evitar posibles sesgos de variables omitidas en nuestra primera etapa, como la educación de los padres agregada a nivel de colegio, es importan hacer que las variables no dicotómicas tengan media cero, ya que esto permite que la correlación entre las variables familiares y las variables agregadas sea igual a cero y de esta manera calculamos con mayor precisión la porción de la varianza explicada por barrios y colegios.

A la hora de centrar variables con una estructura de datos anidados existen dos opciones, centrar las variables respecto a la media general de los datos o centrar las variables respecto a la media del grupo o nivel de anidación. Nosotros elegimos centrar las variables respecto a la media general de los datos (respecto a la media general de la muestra si hablamos de la base muestral, y respecto a la media general de la poblacional sin datos atípicos si hablamos de esta base de datos), esto se debe principalmente a dos motivos: 1) Al tener dos niveles de anidación la decisión de centrarlos datos respecto a la media de alguno de estos dos niveles no es clara y puede sesgar los resultados al usar nuestro modelo completo que tiene en cuenta dos niveles de agrupación, barrios y colegios, y 2) según Hofman y Mark (1998) el centrar respecto a la media general de la muestra permite hacer un mejor cálculo de la varianza y por lo tanto de los efectos aleatorios, esto es sumamente importante si tenemos en cuenta que

nuestro método de estimación funciona en dos etapas y por lo tanto necesitamos un cálculo lo más preciso posible de los efectos aleatorios.

La literatura dice que hay características de los colegios que impactan el aprendizaje, estas variables a nivel de colegio son tomadas del censo educativo c600, entre estas tenemos la educación de los profesores, que esta especificada como la proporción de profesores a nivel de colegio que alcanzaron cada uno de los niveles educativos diferentes (por ejemplo la proporción de profesores que tiene postgrado en el colegio; García, Maldonado, Perry, Rodrigues y Saavedra, 2014). También construimos un índice que tiene en cuenta si la jornada del colegio es ordinaria , la naturaleza del colegio (público o privado), y una medida de la pensión del colegio, el cual es válido según la prueba de Alpha de Cobranch; este índice lo construimos por problemas de multicolinealidad en la segunda etapa. Por otra parte tenemos distintas especificaciones donde incluimos la educación de los padres agregada, el resultado en la prueba agregada (también el rezago de las pruebas en matemáticas y lenguaje del año 2006) y el cambio de variaciones en el resultado de la prueba (matemáticas o lenguaje) y de la edad de los estudiantes a nivel de colegio (Hoxby y Weingarth, 2006). La descripción de estos datos se encuentra en el anexo 3.

Por último tenemos variables a nivel de barrio tomadas del censo poblacional del 2005, entre las cuales se encuentran, la cantidad de universitarios a nivel de UPZ, Goux y Maurin, (2003) muestran que la presencia de personas con diploma universitario es un factor determinante del efecto vecindario sobre la educación. Además de las variables anteriores, tenemos a nivel de barrio datos de homicidios agregados por UPZ, tomados de los datos anuales de la policía nacional (Harding, 2009; Formisano, 2002).

También incluimos un índice que nos muestra la presencia local de clúster a nivel de UPZ de altos resultados, de bajos resultados, y de UPZ de altos resultados rodeadas de UPZ de bajos resultados y viceversa, estas variables son dicotómicas y se construyeron usando el "Local Morans Index", esto lo hacemos con el ánimo de identificar si existe un efecto espacial sobre la educación, lo cual cobra valides si, como mencionamos anteriormente, tenemos en cuenta que existe segregación a gran escala entre el sur y el norte de la ciudad pues es importante identificar si el individuo se encuentra en una zona que grupa buenas o malas características barriales. Aliaga y Álvarez (2010) también encuentran que a una escala más pequeña Bogotá presenta otra estructura de segregación con un patrón menos claro, este fenómeno también se está controlando al incluir el "Local Morans Index", pues este índice tiene en cuenta cada barrio o UPZ en comparación son sus vecinos. La descripción de estos datos esta en el anexo 4 y en el anexo 4.1.

1.4 Estructura de los datos:

Como se verá más adelante, nuestros ejercicios empíricos tendrán en cuenta la agrupación de individuos en barrios y colegios. Nuestra base de datos se compone de 32692 estudiantes, los cuales se encuentran contenidos en 110 UPZ (barrios) de las 117 UPZ en Bogotá y 1181 colegios (esto es teniendo en cuenta cada sede por separado), no sobra aclarar que en la mayoría de estudios sobre educación que tienen en cuenta las condiciones del barrio, los estudiantes viven en el mismo barrio del colegio, por otro lado nuestro estudio tiene tanto estudiantes que viven en el mismo barrio del colegio como estudiantes que no; idealmente el tener esta estructura de datos permite medir con mayor rigurosidad el efecto del vecindario.

En nuestros ejercicios no solo va a importar la pertenencia a un barrio o a un colegio, sino también el traslape entre estos dos grupos. En otras palabras, dado que no todos los estudiantes de un colegio viven en la misma UPZ y viceversa, se generan múltiples grupos de estudiantes que asisten a la misma escuela y viven en el mismo barrio, es decir, nuestros modelos generan un nuevo nivel de anidación para los estudiantes que asisten al mismo colegio y viven en el mismo barrio, donde no necesariamente el colegio queda en el mismo barrio donde habitan. Para ser más específicos, en nuestros datos existen 9571 cruces entre barrios y colegios (que de ahora en adelante denominaremos celdas), las cuales tienen en promedio 15 estudiantes (mínimo tienen un estudiante y máximo 117), de igual manera la desviación típica es de 18.6 estudiantes. Para una mejor descripción de la concentración de barrios por colegio y viceversa, ver el anexo 6.

El proceso para elegir la medida espacial que se va a tomar como barrio se basó en la convergencia del modelo de efectos cruzados (tema en el que ahondaremos en el anexo 5), en otras palabras, si se elige una medida espacial más pequeña de UPZ, el número de grupos se aumenta significativamente y por lo tanto el número de celdas; esto implica que el modelo va a tener que calcular un mayor número de efectos aleatorios con grupos más pequeños y una muestra más desbalanceada, haciendo más difícil la convergencia del modelo. Se puede pensar que una medida espacial más pequeña podría ser más adecuada para desentrañar el efecto de las interacciones sociales a nivel de barrio, pero es difícil definir el tamaño del "hábitat" social de un individuo; además de lo anterior, el nivel de UPZ es más adecuado para controlar por las dotaciones físicas del barrio, esto se debe a que la asignación de recursos públicos, políticas educativas y otros elementos del ordenamiento territorial se organizan a nivel de UPZ[1]. Más específicamente, si tomamos como medida de barrio la UPZ estamos controlando de manera más estricta el efecto que puedan tener las dotaciones físicas del "barrio" del individuo sobre la educación, pues la alcaldía asigna los parques, planes de equipamientos educativos y culturales entre otros, a nivel de UPZ.

[1] Lo anterior está legislado en el decreto 190 del 2004, más específicamente en artículo 46 de ese documento.

2. Primeras estimaciones: modelos con un nivel de anidación.

2.1 Apuntes sobre los modelos jerárquicos multi nivel:

Antes de continuar con nuestros análisis preliminares es importante revisar un poco la teoría de los modelos multi nivel, para tener claridad sobre cómo se debe hacer la comparación con los modelos panel que son nuestro referente (en el anexo 1 realizamos esta comparación con mayor profundidad, esta es importante para darle sustento a la especificación). Los modelos multi nivel se estiman por medio de máxima o cuasi máxima verosimilitud, nosotros preferimos la estimación por cuasi máxima verosimilitud. A pesar de que ambos métodos de estimación no son muy distintos a la hora de encontrar los parámetros fijos (por ejemplo el estimador de la educación de los padres), son distintos a la hora de estimar los efectos aleatorios; pues mientras el estimador por cuasi máxima verosimilitud tiene en cuenta los grados de libertad perdidos por los parámetros a estimar (variables explicativas) el método de máxima verosimilitud no. Esto hace que el método de máxima verosimilitud tenga un sesgo hacia el límite inferior al estimar los efectos aleatorios (Snijders y Bosker, 1999; P. 56).

El objetivo de estos modelos es tener en cuenta la anidación de los datos en grupos para descomponer la varianza de la variable dependiente en los distintos niveles de agregación, esto permite saber que porción de la varianza no explicada por las covariantes, se encuentra explicada por cada uno de los niveles de agregación. Para ilustrar lo anterior usaremos de ejemplo la estructura de nuestro modelo objetivo. Los modelos jerárquicos de efectos cruzados, con variables explicativas al primer nivel tienen la siguiente forma:

Ecuación 1:

$$Y_{ijk} = \gamma_{0jk} + \gamma_{1jk}\Omega_{ijk} + \varepsilon_{ijk}$$

Dónde:

Ecuación 2:

$$\gamma_{0jk} = \alpha_0 + b_{00j} + c_{00k} + v_{0jk}$$

Ecuación 3:

$$\gamma_{1jk} = \alpha_1 + b_{10j} + c_{10k} + v_{1jk}$$

Reemplazando:

Ecuación 4:

$$Y_{ijk} = \alpha_0 + (\alpha_1 + b_{10j} + c_{10k} + v_{1jk})\Omega_{ijk} + b_{00j} + c_{00k} + v_{0jk} + \varepsilon_{ijk}$$

La primera ecuación nos muestra cómo sería la estimación por OLS. La segunda ecuación nos muestra la descomposición de la varianza del intercepto de la primera ecuación, en los efectos de barrio y colegio, el cruce especifico de barrio y colegio en el que se encuentra el individuo (el efecto de la celda anteriormente explicado). La última ecuación nos muestra nuestro modelo a estimar. Si pensamos en un modelo de panel, la ecuación 1 seria nuestro modelo "within" y la ecuación 2 seria nuestro modelo "between". La gran diferencia con un modelo de panel seria que en los modelos panel solo se puede tener en cuenta un único nivel de agregación para estimar el modelo "between" (Leeden, Busing y Meijer, 1997). La ecuación 3 nos muestra la descomposición de la varianza del estimador asociado a las variables familiares en la primera ecuación, según la anidación de los datos en barrios y colegios y la celda a la que pertenece el individuo; lo anterior implica que el efecto de las variables familiares se puede desviar del efecto promedio según el barrio, el colegio o el cruce especifico barrio-colegio del individuo, esta es otra de las diferencias de estos modelos con los modelos panel.

Para este modelo Y_{ijk} (la variable dependiente) representa el resultado en la prueba Saber11 en matemáticas o lenguaje, en donde los subíndices i se refieren al individuo, j al barrio y k al colegio. α_0 es el intercepto, c_{00k} es el efecto aleatorio asociado al nivel de colegio, b_{00j} es el efecto aleatorio a nivel de barrio, y v_{0jk} es la interacción del efecto aleatorio asociado a asistir al colegio j y vivir en el barrio k (la variación intra-celda). Estos modelo sufren de problemas de convergencia a medida que se tiene un mayor número de grupos no balanceados, lo anterior se refleja particularmente en el hecho que v_{0jk} va a ser un conjunto de efectos aleatorios por cada cruce de barrio y de colegio presente en la muestra o celda como lo definimos anterior mente. Por otra parte, ε_{ijk} es el término de error asociado a todo modelo econométrico, y γ_{0jk} sería la constante asociada a un modelo OLS ("naive").

El término Ω_{ijk} es un vector que incluye todas las variables familiares, por otra parte el término α_1 es el efecto promedio (o estimador promedio) de las variables a nivel de familia. Respecto a la parte aleatoria del modelo, el término b_{10j} indica las perturbaciones del efecto promedio de las variables familiares a nivel de barrio, el término c_{10k} indica las perturbaciones del efecto promedio de las variables familiares a nivel de colegio, y el término v_{1jk} indica las perturbaciones del efecto promedio de las variables familiares a nivel de celda; si cualquiera de los tres términos anteriores es distinto a cero, implica que este modelo tiene pendientes aleatorias referentes a las condiciones familiares.

Nuestro método permitirá obtener los efectos aleatorios asociados al nivel de barrio y de colegio, eliminando la parte de la varianza de la variable dependiente explicada por las condiciones familiares (observables). Para los estimadores de las variables de familia se permitirá la presencia de efectos aleatorios a nivel de colegio que hasta el momento es el nivel de anidación que más explica la varianza de la variable dependiente, el no tener en cuenta estos cambios en las pendientes podría llevar a no estimar correctamente la varianza del modelo.

No sobra aclarar que en esta especificación el intercepto y la pendiente son aleatorias; la interpretación de lo anterior, como lo muestran Albright y Marinova (2010), es básicamente, que cada institución educativa o cada UPZ puede tener un efecto fijo o pendiente distinta, de igual manera en este tipo de regresiones existe una constante general para toda la muestra. Lo anterior tiene mucho sentido si tenemos en cuenta que los distintos grupos no son replicas independientes de la misma estructura, en otras palabras, estos modelos nos permiten suponer que tanto los barrios como los colegios pueden tener un efecto distinto sobre la educación y una pendiente distinta para el efecto de las variables familiares (esto se llama heterogeneidad de las regresiones).

2.2 Modelos vacíos de un nivel:

A continuación comparamos nuestros modelos panel de efectos fijos y aleatorios sin variables explicativas con nuestros modelos jerárquicos de un nivel, lo importante es comparar las distintas estimaciones de los efectos entre grupos y al interior de los grupos. En las siguientes tablas mostramos nuestros resultados en matemáticas y lenguaje para barrios y colegios.

Tabla 3:

Comparación de modelos, barrio y colegio (Panel de efectos fijos, aleatorios y jerárquicos de un nivel):

Comparación Efectos fijos, efectos aleatorios, multinivel: colegios

Muestral

	Lenguaje	Matemáticas	Lenguaje	Matemáticas	Lenguaje	Matemáticas	Lenguaje	Matemáticas
con sexo y edad	Efectos aleatorios		Efectos fijos		multi nivel		Cuasi Máxima verosimilitud	
between	3.842	4.766	4.445	5.563	3.774	4.967	3.774	4.967
within	6.135	7.661	6.135	7.661	6.136	7.659	6.136	7.659

vacío	Efectos aleatorios		Efectos fijos		multi nivel		Cuasi Máxima verosimilitud	
between	3.852	4.752	4.406	5.47	3.699	4.853	3.699	4.853
within	6.172	7.818	6.172	7.818	6.174	7.815	6.174	7.815

Comparación Efectos fijos,Efectos aleatorios, multinivel: barrios

Muestral

	Lenguaje	Matemáticas	Lenguaje	Matemáticas	Lenguaje	Matemáticas	Lenguaje	Matemáticas
vacío	Efectos aleatorios		Efectos fijos		multi level		Cuasi Máxima verosimilitud	
between	2.935	3.848	3.13	4.102	2.97	3.86	2.97	3.86

with in	6.689	8.737	6.689	8.737	6.689	8.737	6.689	8.737

con sexo y edad	Efectos aleatorios		Efectos fijos		multi level		Cuasi Máxima verosimilitud	
between	2.468	3.134	3.245	4.218	3.097	3.981	3.097	3.981
with in	6.656	8.596	6.656	8.596	6.656	8.596	6.656	8.596

En las tablas anteriores se pueden notar que las diferencias entre los modelos panel con efectos aleatorios y los modelos jerárquicos de un nivel no son muy grandes, pero las diferencias entre modelos jerárquicos de un nivel y los modelos con efectos fijos si lo son; según Snijders y Bosker (1999), estas diferencias entre efectos aleatorios pueden generarse por lo desbalanceado de los distintos grupos de colegios y barrios (en otras palabras, existen grupos con muy pocos individuos y grupos con más de 100 individuos); a la hora de tratar este problema los modelos con efectos aleatorios son más adecuados que los modelos con efectos fijos, puesto que los modelos de efectos fijos, estiman cada uno de los errores aparte, entonces los coeficientes serán sobre estimados en los casos de los grupos pequeños, por falta de información, esto hace que los parámetros tengan grandes errores standart. Por otra parte los efectos aleatorios tienen en cuenta el supuesto de que los efectos de grupo son independientes e idénticamente distribuidos, en otras palabras, para el cálculo tienen en cuenta la distribución de los datos; esto permite contrarrestar la escasez de casos, haciendo la inferencia más precisa.

2.3 Prueba de anidación:

Antes de continuar con la estimación de nuestros modelos a un nivel, debemos poner a prueba que el pertenecer a un grupo si tiene un efecto sobre los resultados en las pruebas saber 11, esto lo realizamos comparando un modelo vacío con un modelo vacío anidado por medio de un test de tasa de verosimilitud ("Likehood-ratio test"), más específicamente estamos comparando el logaritmo de la función de verosimilitud maximizada de un modelo con anidación y uno sin anidación (Rabe-Hesketh Skrondal, 2008).

Prueba de única anidación: matemática barrios			
Likelihood-ratio test	LR chi2(1)	=	4318.68

(Supuesto: modelo anidado en barrios barrio no contiene el modelo sin anidación)	Prob > chi2	=	0

Prueba de única anidación: lenguaje barrios

Likelihood-ratio test	LR chi2(1)	=	4445.35
(Supuesto: modelo anidado en barrios barrio no contiene el modelo sin anidación)	Prob > chi2	=	0

Prueba de única anidación: matemática colegios

Likelihood-ratio test	LR chi2(1)	=	9461.1
(Supuesto: modelo anidado en colegios no contiene el modelo sin anidación)	Prob > chi2	=	0

Prueba de única anidación: lenguaje colegios

Likelihood-ratio test	LR chi2(1)	=	7618.05

(Supuesto: modelo anidado en colegios no contiene el modelo sin anidación)	Prob > chi2 =	0

A partir de la tabla anterior podemos decir que, tanto para matemáticas como para lenguaje la anidación en barrios y colegios si explica parte de la varianza de la variable dependiente, se podría pensar que estos modelos jerárquicos de un nivel contienen a los modelos sin anidación. Nuestro paso a seguir es estimar modelos jerárquicos de un solo nivel con variables explicativas a nivel familiar y personal. Sobre la especificación de esta primera etapa se hicieron las pruebas correspondientes y no se encontró evidencia de efectos no lineales, en este aspecto Cook et al (2002) encontró que tanto pares, colegios, barrios y familia eran significativos, pero que los efectos de estas variables eran aditivos.

2.4: Regresiones con variables explicativas y efectos aleatorios, modelos de un nivel.

Tabla4:

modelo a un nivel	colegios			
	Muestral		Poblacional sin datos atípicos	
VARIABLES	lenguaje	matemática	lenguaje	matemática
Sexo	-0.0341	2.804***	-0.117**	2.683***
	(0.0756)	(0.0949)	(0.0476)	(0.0584)
Edad centrada del estudiante	-0.518***	-0.715***	-0.524***	-0.697***
	(0.0331)	(0.0416)	(0.0205)	(0.0252)
Años de educación de la madre centrados	0.103***	0.119***	0.0845***	0.0929***
	(0.0130)	(0.0163)	(0.00804)	(0.00987)
Años de educación del padre centrados	0.145***	0.110***	0.117***	0.0907***
	(0.0122)	(0.0153)	(0.00756)	(0.00928)
	-0.772***	-0.613***	-0.731***	-0.586***

Hermano desertor menor de 15 años	(0.179)	(0.224)	(0.105)	(0.129)
Sin hermanos	-0.0318 (0.0750)	-0.201** (0.0939)	-0.0280 (0.0485)	-0.218*** (0.0594)
Ingreso familiar centrado	0.507*** (0.0344)	0.595*** (0.0435)	0.420*** (0.0234)	0.473*** (0.0290)
Controles laborales	si	si	si	si
Constante	48.24*** (0.161)	46.03*** (0.212)	47.98*** (0.110)	45.84*** (0.148)
Observations	32,692	32,692	79,231	79,231
Numero de Cedes	1,181	1,181	1,441	1,441
varianza entre colegios (between)	2.660659	3.984474	2.867954	4.145889
varianza al interior (within)	6.109829	7.636909	6.123828	7.505056
*** p<0.01, ** p<0.05, * p<0.1				

En las regresiones anteriores podemos ver que para casi todas las variables, tanto el signo como la significancia de las mismas es igual si se usa la muestra poblacional sin datos atípicos que si se usa la muestra final. Entrando más en detalle podemos ver que nuestra regresión replica algunas de las evidencias empíricas halladas por otros autores (Mediavilla y Martinez, 2009; Ready, 2010) más detalladamente podemos ver que el ser hombre tiene una relación positiva con las pruebas de matemáticas, al comparar el resultado de ser hombre en la prueba de lenguaje respecto a una regresión "naive" (ver anexo 1 regresiones preliminares) podemos ver que deja de ser significativo, es posible que esto se deba a una mejor especificación de la varianza del modelo; también podemos ver que los estudiantes en extra edad y/o que tienen hermanos desertores (menores de 15 años) tienden a tener peores resultados en ambas pruebas; por ultimo podemos ver que a mayor educación e ingreso de los padres mejores resultados en las pruebas saber 11.

Ahora analizaremos nuestros resultados teniendo en cuenta la anidación en barrios:

Tabla5:

modelo a un nivel	barrios Muestral	
VARIABLES	lenguaje	Matemática
sexo	-0.222***	2.660***
	(0.0718)	(0.0931)
Edad centrada del estudiante	-0.462***	-0.585***
	(0.0328)	(0.0425)
Años de educación de la madre centrados	0.154***	0.194***
	(0.0133)	(0.0173)
Años de educación del padre centrados	0.184***	0.170***
	(0.0125)	(0.0163)
Hermano desertor menor de 15 años	-0.944***	-0.808***
	(0.185)	(0.240)
Sin hermanos	0.0292	-0.189*
	(0.0772)	(0.100)
Ingreso familiar centrado	0.780***	1.077***
	(0.0334)	(0.0433)
Controles laborales	si	si
Constante	48.67***	46.59***
	(0.203)	(0.277)
Observaciones	32,692	32,692
Numero de UPZ	110	110
varianza entre barrios (between)	1.444605	2.043004
varianza al interior del barrio (within)	6.427974	8.320136
*** p<0.01, ** p<0.05, * p<0.1		

Los resultados de estas regresiones que controlan por variaciones barriales son similares a los encontrados en los modelos anteriores, en otras palabras, la educación de los padres y su ingreso tienen un efecto positivo; y el estar en extra edad y tener hermanos desertores tiene un efecto negativo. Por otra parte el efecto del género es el mismo, positivo para hombres en la prueba de matemáticas. Sobre el género en la prueba de lenguaje, al comparar nuestro modelo jerárquico de un nivel anidado en colegios con nuestro modelo jerárquico de un nivel anidado en barrios y con el modelo OLS ("naive", ver anexo 1) podemos ver que desaparece el efecto positivo de ser mujer en las pruebas de lenguaje, es muy probable que esto se deba a no tener en cuenta el efecto de la anidación de los datos en barrios o colegios, pues según Snijders y Bosker (1999) el omitir este tipo de estructuras aumenta el error de tipo 1.

3. Regresiones con variables explicativas y efectos aleatorios cruzados, modelos multi nivel:

3.1 Prueba de doble anidación:

Estos modelos son el objetivo central de nuestro trabajo, el cual es lograr descomponer la varianza no explicada del modelo en los dos grupos de anidación, barrios y colegios. Como ya mostramos anteriormente es válido usar este tipo de modelos con efectos aleatorios para tratar este problema, por otra parte, las pruebas anteriores eran de suma importancia pues la prueba de tasa de verosimilitud ("Likehood-ratio test") de la que hablamos anteriormente no se puede aplicar con toda libertad a modelos con más de un parámetro aleatorio. Lo anterior se debe a que se desconoce con exactitud el número de efectos aleatorios que deben ser restringidos a cero, en especial si tenemos en cuenta la existencia de covarianzas; de igual manera si se rechaza la hipótesis nula (que dice que los efectos aleatorios son iguales a cero) se está encontrando un límite superior, por lo tanto si se rechaza la hipótesis nula con los grados de libertad propuestos implica que se rechaza esta hipótesis con los grados de libertad que en realidad tiene el modelo.

Dado lo anterior, para estar totalmente seguros de la doble anidación que proponemos como modelo ideal, podemos comparar un modelo con un único nivel de anidación, con un modelo que posee doble anidación, es decir estaríamos comparando si un modelo que solo tiene en cuenta la anidación en colegios estaría contenido por un modelo anidado en colegios y barrios, de esta manera reducimos el número de parámetros a ser comparados, de todas maneras el estimado continua siendo un límite superior pero de esta forma no estará tan lejano:

Prueba de doble anidación: matemáticas colegios			
Likelihood-ratio test	LR chi2(1)	=	18.01
(Hipótesis: modelo vacío multi nivel no contiene modelo anidado en solo colegios)	Prob > chi2	=	0

Prueba de doble anidación: lenguaje colegios			
Likelihood-ratio test	LR chi2(1)	=	71.68
(Hipótesis: modelo vacio multi nivel anidación no contiene modelo anidado en solo colegios)	Prob > chi2	=	0

Prueba de doble anidación: matemática barrios			
Likelihood-ratio test	LR chi2(1)	=	5177.88
(Hipótesis: modelo vacío multi nivel de no contiene modelo anidado en solo colegios)	Prob > chi2	=	0

Prueba de doble anidación: lenguaje barrios			
Likelihood-ratio test	LR chi2(1)	=	3249.34
(Hipótesis: modelo vacío multi nivel de no contiene el modelo anidado en solo barrios)	Prob > chi2	=	0

Las pruebas anteriores nos muestran que la hipótesis de doble anidación es válida en todos los casos. Este procedimiento también puede ser usado para saber si la celda (los grupos generados por el cruce de barrios y colegios) explica efectivamente algo de la varianza de la variable dependiente, en particular este último nivel de anidación es sumamente interesante, pues si existe un sesgo de selección que relacione la pertenencia a estos dos grupos estaría contenido en la varianza explicada por la celda, pero en si mismo controla por la con junción

de las condiciones del barrio y el colegio. A continuación mostramos los resultados de esta prueba:

Prueba sobre la celda: Matemáticas.			
Likelihood-ratio test	LR chi2(1)	=	7.63
(Supuesto: Modelo multinivel con celda contiene a modelo multi nivel sin celda)	Prob > chi2	=	0.0057

Prueba sobre la celda: Lenguaje.			
Likelihood-ratio test	LR chi2(1)	=	3.74
(Supuesto: Modelo multinivel con celda contiene a modelo multi nivel sin celda)	Prob > chi2	=	0.0531

En este caso también podemos decir que la celda es válida tanto para matemáticas como para lenguaje. De igual manera en el caso de la regresión de lenguaje esta prueba se encuentra justo en el límite. Dados los resultados anteriores podemos continuar con la estimación de nuestro modelo ideal.

3.2 Resultados de los modelos con doble anidación:

En los siguientes modelos tenemos dos grupos en lo que respecta a la estimación de efectos aleatorios. Modelos que solo suponen un componente aleatorio en el intercepto, en barrios, colegios y el cruce de estos dos grupos (las celdas), y modelos que además de tener estos componentes aleatorios suponen que el efecto de la educación de los padres puede variar entre colegios, en resumidas cuentas, hay un componente asociado al estimador de la educación del padre y de la madre, que varía según colegios; esto lo suponemos porque en algunos estudios se ha mostrado que el efecto de las condiciones contextuales del individuo (Patacchini y Zenou, 2009; Ready 2010), dependen de la actitud de los padres respecto al aprendizaje del hijo, dado que no tenemos esta variable, podemos modelar que si pertenecer

a un colegio dado, potencia o reduce el efecto de la educación de los padres sobre los resultados educacionales de su hijo.

Tabla 6:

Modelo con doble anidación, prueba de lenguaje.

VARIABLES	vacío lenguaje	vacío, pendiente lenguaje	constante lenguaje	pendiente lenguaje
Sexo			-0.0464 (0.0755)	-0.0399 (0.0755)
Edad centrada del estudiante			-0.526*** (0.0331)	-0.525*** (0.0331)
Años de educación de la madre centrados			0.108*** (0.0138)	0.100*** (0.0130)
Años de educación del padre centrados			0.152*** (0.0130)	0.141*** (0.0122)
Hermano desertor menor de 15 años			-0.754*** (0.179)	-0.767*** (0.179)
Sin hermanos			-0.0274 (0.0751)	-0.0333 (0.0751)
Ingreso familiar centrado			0.492*** (0.0347)	0.489*** (0.0347)
Controles laborales			Si	Si
Constante	48.08*** (0.146)	47.99*** (0.146)	48.38*** (0.169)	48.43*** (0.171)
UPZ	0.9668244	0.8846841	0.5754326	0.560751
Colegio	3.211632	2.814337	2.511128	2.345096
Celda	0.5579002	0.5029559	0.4112985	0.4111659
Error	6.15945	6.121486	6.104554	6.089918

educación padre		0.2210713		0.1200715
educación madre		0.1943128		0.1155943
Observaciones	32,692	32,692	32,692	32,693
Standard errors in parentheses	*** p<0.01, ** p<0.05, * p<0.1			

El hecho de ser hombre no tiene efecto sobre las pruebas de lenguaje, esto puede implicar que el género deja de ser importante en el aspecto del lenguaje, por otra parte, no solo este estudio ha encontrado que ser mujer se correlaciona con peores resultados educativos, entre otros está el estudio de Gabiria y Barrientos (2001) quienes muestran que ser hombre se correlaciona con mejores resultados tanto en el puntaje general como en la prueba de matemáticas y lenguaje. De igual manera en varios de nuestros modelos estimados anteriormente el ser mujer tenía efectos positivos sobre la prueba de lenguaje, es posible que este cambio en la significancia se deba a una correcta estimación de la varianza del modelo. Además este resultado indica que persisten ciertas estructuras patriarcales en Bogotá, en este aspecto Domínguez (2004) argumenta que no solo existen diferencias reales en las instituciones educativas (como normas o cursos especializados para cada género), sino que además persisten diferencias en las expectativas educativas por género.

Adentrándonos en las variables familiares, la edad del estudiante es una proxi de su desempeño académico, pues es muy posible que estudiantes en extra edad hayan reprobado algún curso, y como es de esperarse su signo es negativo. El hecho de tener un hermano desertor muestra que, por un lado la familia puede enfrentar dificultades económicas, las cuales se relacionan con peores resultados educativos (Brook-Gun et al, 2007), y por otro lado la familia valora menos la educación, condición que a su vez se relaciona con un peor desempeño (Patacchini y Zenou, 2009). Por otra parte padres más educados y con mayores ingresos, se correlacionan con mejores resultados en las pruebas. De igual manera hay que tener cuidado con los canales de trasmisión, pues el efecto de la educación puede ser o un efecto directo de aprendizaje o un efecto de la valoración de la familia por la educación. Sobre el ingreso, el canal de transmisión tampoco es claro, según Gabiria y Barrientos (2001) puede ser un efecto de elección del colegio, pero el efecto de las condiciones materiales no debe ser despreciable.

Sobre los efectos aleatorios, como era de esperarse, en todas las especificaciones, el colegio explica una porción de la varianza mucho mayor que el barrio, más específicamente para nuestro modelo completo, mientras el barrio explica casi un 8.95% el colegio aproximadamente un 30% y la celda cerca de un 5.1% de la varianza de la variable dependiente. Además podemos ver como la porción de la varianza explicada por colegios y

barrios baja al incluir las covariantes familiares. También es muy interesante que al incluir un efecto aleatorio de la educación de los padres a nivel de colegio, el efecto del colegio se reduce en una medida no despreciable, esto indica que no solo importan las condiciones del colegio, sino que además el efecto del colegio se puede ver potenciado o reducido según la educación de los padres, esto puede ser producto de la valoración de la educación a nivel familiar (mas el efecto de estudiar o ser educado en parte por los padres) y de que ciertos colegios tengan programas de padres de familia.

Tabla 7:

Modelo con doble anidación, prueba de matemáticas.

VARIABLES	vacio matemática	vacio: pendiente matemática	constante matemática	pendiente matemática
sexo			2.802***	2.658***
			(0.0948)	(0.0944)
Edad centrada del estudiante			-0.717***	-0.358***
			(0.0416)	(0.0423)
Años de educación de la madre centrados			0.118***	0.232***
			(0.0163)	(0.0174)
Años de educación del padre centrados			0.109***	0.209***
			(0.0153)	(0.0162)
Hermano desertor menor de 15 años			-0.612***	-0.845***
			(0.224)	(0.243)
Sin hermanos			-0.205**	-0.156
			(0.0939)	(0.102)
Ingreso familiar centrado.			0.592***	1.491***
			(0.0436)	(0.0408)
Controles laborales			Si	Si
Constante	47.27***	47.18***	46.13***	46.41***

	(0.164)	(0.164)	(0.214)	(0.184)
upz	0.5863806	0.5221777	0.2569291	0.2593112
Colegio	4.671767	4.32467	3.981627	3.915249
Celda	0.8635234	0.7204402	0.6766715	0.6708241
Error	7.780739	7.742304	7.619098	7.613417
educación padre		0.2015449		0.0988929
educación madre		0.2164549		0.0567377
Observaciones	32,692	32,692	32,692	32,692
Standard errors in parentheses	*** p<0.01, ** p<0.05, * p<0.1			

Como muestra la literatura, a los hombres les va estadísticamente mejor en matemáticas, incluso al controlar por variables familiares y por la anidación de los individuos en barrios y colegios. La edad continúa siendo una proxy de rendimiento académico. Por otra parte el resto de variables familiares, como la educación de los padres o el ingreso, tienen el mismo signo que en nuestras regresiones anteriores y magnitudes similares.

Al igual que en la prueba de lenguaje, para todas las especificaciones, la varianza explicada por el colegio es mayor, pero en este caso el efecto del barrio es radicalmente menor, más específicamente, para el modelo vacío, el barrio explica aproximadamente un 4.3% de la varianza, mientras el colegio explica cerca de un 34% y la celda un 6.2%; esto puede indicar que la transmisión del aprendizaje del lenguaje se da más por interacciones sociales (de igual manera la escuela sigue siendo lo más importante), y el efecto sobre matemáticas sea más una cuestión de calidad educativa. Dado que en la estructura educacional en Colombia el barrio de los estudiantes y el de los colegios no se superpone perfectamente, es importante comparar el efecto celda con el efecto del barrio; en el caso de la prueba de lenguaje el efecto aleatorio de la UPZ es mayor que el efecto de la celda, en cambio en la prueba de matemáticas sucede lo contrario. Esto es evidencia a favor de que el aprendizaje del lenguaje pueda depender más del contexto social del individuo, mientras que en matemáticas, a pesar de que el barrio importa, es mucho más importante la escuela, y la combinación barrio-colegio; mientras en lenguaje es más importante el barrio que la combinación barrio-colegio. Este resultado ha sido mencionado por algunos autores (Steele Vignoles y Jenkins 2007).

Otro resultado interesante es que el efecto de la educación de los padres (el estimador) cambia según la pertenencia al colegio, más específicamente, el impacto que tiene la educación de los padres sobre los resultados en las pruebas saber 11 cambia entre los distintos colegios (distintas pendientes). De igual manera la educación de la madre tiende a ser más estable

entre grupos que la educación del padre, al controlar por variables familiares, lo anterior podría indicar que las madres tienden a estar más involucradas en la educación de sus hijos en promedio, en cambio en los padres hay más heterogeneidad en el efecto. Para complementar lo anterior, al extraer de la base poblacional aquellos estudiantes que no conocen la educación de sus padres, solo 770 no conocen la educación de su madre y en cambio 7778 no conocen la educación de su padre. Esto puede indicar que el abandono familiar es más común por parte del padre que de la madre, pero no solo esto, sino que nos podría llevar a pensar que es posible que la tarea de educar y estar pendiente de los hijos sea una tarea más femenina, lo que nos lleva a pensar en una estructura familiar patriarcal.

3.3 Apuntes sobre la segunda etapa.

A partir de los efectos aleatorios anteriormente estimados, tendremos las variables dependientes de nuestro segundo paso de estimación. Este paso consiste en encontrar las variables más correlacionadas a barrio y colegio. Específicamente, estimaremos el efecto aleatorio asociado al vecindario, con las variables del vecindario, y el efecto aleatorio del colegio con las variables de colegio (modelo de intercepto aleatorio). Con estas estimaciones podremos saber qué variables son significativas a qué nivel, pero lo más importante, podremos comparar la significancia de los estimadores de las regresiones que usan los efectos aleatorios del modelo vacío, contra los modelos que usan los efectos aleatorios del modelo con variables familiares. En este punto los modelos a estimar serían los siguientes:

$$\hat{b_{00j}} = A_{00j} + A_{01j}X_j + E_j$$

$$\hat{c_{00k}} = A_{00k} + A_{01k}Z_k + E_k$$

En este caso la primera y segunda ecuación son nuestros modelos a estimar por OLS, que buscan explicar los efectos aleatorios a nivel de barrio y de colegio respectivamente; estos son los efectos aleatorios calculados para el intercepto del modelo con variables familiares.

En estas ecuaciones, X_j y Z_k son vectores de variables a nivel barrio y colegio respectivamente. Los términos A_{00j} y A_{00k} indican el intercepto de las regresiones; los términos A_{01j} y A_{01k} indican la pendiente de las regresiones (el efecto de las condiciones de barrio y colegio respectivamente sobre los efectos aleatorios); y por último, E_j indican el error asociado a esta estimación.

Las estimaciones a continuación utilizan clúster de errores (revisar nombre del estimador sanduich…), las estimaciones de colegio tienen en cuenta el clúster de errores a nivel de barrio y las estimaciones de barrio tienen en cuenta el clúster a nivel de colegios.

3.4 Resultados de la segunda etapa, vecindarios.

Tabla 8:

Segundas etapas, Prueba de matemáticas barrios:

Para matemáticas:

VARIABLES	prueba sobre la variable dependiente matemática	modelo vacío Efecto aleatorio del barrio: matemática	Constante Efecto aleatorio del barrio: matemática	pendiente Efecto aleatorio del barrio: matemática
Proporción de universitarios por UPZ	16.23*** (1.209)	1.717*** (0.0498)	0.325*** (0.0210)	0.336*** (0.0217)
homicidios totales en la UPZ	-0.00198* (0.00106)	-0.00096*** (8.16e-05)	-5.53e-05** (2.20e-05)	-6.50e-05*** (2.27e-05)
Clúster de UPZ con altos resultados rodeadas por UPZ de altos resultados.	3.721*** (0.741)	0.0986*** (0.0164)	0.0561*** (0.00844)	0.0577*** (0.00867)
Clúster de UPZ con bajos resultados rodeadas por UPZ de bajos resultados.	-0.236 (0.300)	-0.0648*** (0.0166)	0.0111** (0.00506)	0.0103** (0.00518)
Clúster de UPZ con altos resultados rodeadas por UPZ de bajos resultados	5.403 (3.823)	-0.485*** (0.0129)	-0.0789*** (0.00546)	-0.0829*** (0.00565)
Clúster de UPZ con bajos resultados rodeadas por UPZ de altos resultados	-0.659 (0.631)	0.0173 (0.0113)	0.0240*** (0.00340)	0.0240*** (0.00350)
Constante	44.94*** (0.220)	-0.268*** (0.0121)	-0.0689*** (0.00468)	-0.0703*** (0.00482)
Observaciones	32,616	32,616	32,616	32,616
R-squared	0.120	0.606	0.267	0.270

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Tabla 9:

Segundas etapas, Prueba de lenguaje barrios:

VARIABLES	prueba sobre la variable dependiente lenguaje	modelo vacío Efecto aleatorio del barrio:	constante Efecto aleatorio del barrio:	Pendiente Efecto aleatorio del barrio:
Proporción de universitarios por UPZ	13.80*** (0.760)	3.822*** (0.0949)	1.675*** (0.0578)	1.671*** (0.0577)
Homicidios totales en la UP	-0.00241*** (0.000814)	-0.000825*** (0.000192)	2.80e-05 (0.000132)	2.56e-05 (0.000130)
Clúster de UPZ con altos resultados rodeadas por UPZ de altos resultados	2.017*** (0.431)	0.311*** (0.0296)	0.378*** (0.0205)	0.351*** (0.0202)
Clúster de UPZ con bajos resultados rodeadas por UPZ de bajos resultados	-0.794*** (0.195)	-0.675*** (0.0295)	-0.308*** (0.0192)	-0.311*** (0.0185)
Clúster de UPZ con altos resultados rodeadas por UPZ de bajos resultados	2.914 (1.932)	-0.792*** (0.0242)	-0.296*** (0.0155)	-0.313*** (0.0154)
Clúster de UPZ con bajos resultados rodeadas por UPZ de altos resultados	-2.592*** (0.946)	0.0753** (0.0322)	0.0566*** (0.0146)	0.0594*** (0.0139)

Constante	46.14*** (0.158)	-0.634*** (0.0228)	-0.342*** (0.0147)	-0.339*** (0.0145)
Observaciones	32,616	32,616	32,616	32,616
R-squared	0.128	0.804	0.675	0.676
Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1				

La primera columna de regresión se corre contra la variable dependiente (ya sea matemáticas o lenguaje) esto se hace con el ánimo de mostrar que el efecto es robusto no solo para explicar los efectos aleatorios, sino también para explicar como tal a la variable dependiente de nuestra primera etapa.

Las regresiones OLS anteriores muestran un efecto positivo del porcentaje de personas con grado universitario en la UPZ para todas las especificaciones, tanto en la prueba de lenguaje como de matemáticas; esto puede ser evidencia a favor de un efecto barrio por medio de un modelo a seguir (Jencks y Mayer, 1990; Wilson, 1987), lo anterior significa que si los estudiantes se encuentran rodeados por mas universitarios tienen más incentivos a estudiar más y a buscar un mayor nivel educativo, lo cual implica mejores resultados.

Por otra parte el crimen muestra tener un efecto negativo sobre los resultados, además explica parte del efecto negativo asociado a los barrios, esto se puede dar por dinámicas de estrés, o de sustitución, pues el crimen se puede presentar como una opción de vida. Otros estudios también han encontrado que el crimen del barrio se traduce en condiciones desfavorables para la educación; por ejemplo Harding (2009) encuentra que el efecto del barrio que se correlaciona con el abandono educativo se explica, para los hombres, en un 65% por la violencia y para las mujeres en un 100%. Además esta variables es importante pues muchos de los homicidios ocurren en la cercanía del hogar del victimario y por lo tanto es un buen indicador de las condiciones del barrio (Formisano, 2002)

Por último nuestra medida de "Local Morans Index" muestra que posiblemente exista un efecto espacial en lo referente al efecto del vecindario, en otras palabras clústeres de UPZ con mayores resultados en promedio se relacionan con efecto vecindario positivo sobre los resultados en la pruebas; en cambio clústeres de barrios con bajos resultados en las pruebas se correlacionan con un peor efecto vecindario. El efecto de vivir en un barrio malo rodeado de barrios buenos no es robusto en las distintas regresiones. De igual manera El hecho de que exista un efecto significativo de la existencia de clústeres puede ser indicio de que no solo importe el barrio del individuo, sino que además importan las condiciones de los barrios circundantes.

Otro resultado que da soporte a una de nuestras hipótesis más interesantes, que el lenguaje depende más del efecto de las interacciones sociales es que, el R-cuadrado de los modelos que explican el efecto vecindario, es mucho mayor para la prueba de lenguaje que para la de matemáticas; esto cobra real importancia si tenemos en cuenta que la variable de porcentaje de universitarios en la UPZ tiene una magnitud mucho mayor en la prueba de lenguaje que en la de matemáticas y además. En contra de esto el efecto de los homicidios deja de ser significativo en las últimas dos especificaciones del efecto aleatorio de la prueba de lenguaje, pero uno de los canales de transmisión del efecto del crimen es que se muestra como un sustituto a la educación, esto hace que este resultado no invalide nuestra hipótesis de que la prueba de lenguaje depende mas del contexto social del individuo.

3.4 Resultados de la segunda etapa, colegios.

Tabla 10:

Segundas etapas, Prueba de matemática, colegios:

	prueba sobre la variable dependiente	modelo vacío	Constante	Pendiente
		Efecto aleatorio del colegio:	Efecto aleatorio del colegio:	Efecto aleatorio del colegio:
VARIABLES	matemática	matemática	matemática	matemática
Cambio de variaciones de la edad por colegio.	-85.76***	-63.15***	-56.47***	-0.0205***
	(4.750)	(4.739)	(4.773)	(0.00604)
Cambio de variaciones de la prueba de matemáticas.	4.862*	3.349	2.735	0.0323***
	(2.887)	(2.454)	(2.507)	(0.00467)
Índice de colegios privados, valor de la pensión y jornada.	0.179***	0.228***	0.191***	-4.26e-05
	(0.0503)	(0.0301)	(0.0348)	(0.000145)
Promedio de educación de los padres	0.587***	0.564***	0.299***	0.000754***
	(0.0401)	(0.0370)	(0.0385)	(9.65e-05)
Porcentaje de profesores con bachillerato pedagógico.	0.173	0.474	0.120	-0.00631
	(0.994)	(0.637)	(0.692)	(0.00431)
Porcentaje de profesores normalista superiores.	0.886	1.156*	1.443*	0.00965***
	(0.845)	(0.680)	(0.750)	(0.00285)

Porcentaje de profesores con estudios técnicos.	2.325** (1.055)	2.376*** (0.879)	2.595*** (0.941)	0.00616* (0.00349)
Porcentaje de profesores con estudios técnicos en pedagogía.	1.524 (0.996)	1.895** (0.777)	1.710* (0.907)	0.00129 (0.00270)
Porcentaje de profesores con estudios profesionales.	2.179*** (0.807)	2.090*** (0.537)	2.232*** (0.620)	0.00233 (0.00330)
Porcentaje de profesores con estudios profesionales en pedagogía.	2.419* (1.294)	2.061** (0.944)	2.499** (1.159)	0.00631 (0.00412)
Porcentaje de profesores con postgrado.	5.094*** (1.197)	5.057*** (1.166)	4.881*** (1.219)	0.0124*** (0.00307)
Porcentaje de profesores con postgrado pedagógico.	3.761*** (0.969)	3.712*** (0.820)	4.498*** (0.876)	0.00421 (0.00328)
Constante	69.84*** (1.290)	14.94*** (1.136)	15.13*** (1.147)	-0.00965*** (0.00269)
Observaciones	32,643	32,643	32,643	32,643
R-squared	0.271	0.701	0.582	0.120
Robust standard errors in parentheses	*** p<0.01, ** p<0.05, * p<0.1			

Tabla 11:

Segundas etapas, Prueba lenguaje colegios:

VARIABLES	prueba sobre la variable dependiente lenguaje	modelo vacío Efecto aleatorio del colegio: lenguaje	Constante Efecto aleatorio del colegio: lenguaje	pendiente Efecto aleatorio del colegio: lenguaje
Cambio de variaciones de la edad por colegio.	-43.36*** (1.902)	-26.54*** (1.704)	-22.08*** (1.719)	-20.45*** (1.480)
Cambio de variaciones de la prueba de lenguaje.	-26.36***	-24.96***	-22.52***	-21.00***

	(2.240)	(1.590)	(1.511)	(1.396)
Índice de colegios privados, valor de la pensión y jornada.	0.120***	0.129***	0.102***	0.101***
	(0.0369)	(0.0230)	(0.0212)	(0.0198)
Promedio de educación de los padres.	0.621***	0.474***	0.237***	0.201***
	(0.0292)	(0.0222)	(0.0208)	(0.0178)
Porcentaje de profesores normalista superiores.	-0.693	-0.140	-0.305	-0.149
	(0.785)	(0.608)	(0.575)	(0.538)
Porcentaje de profesores con estudios técnicos.	-1.494**	-0.825	-0.882*	-0.969**
	(0.694)	(0.529)	(0.513)	(0.465)
Porcentaje de profesores con estudios técnicos en pedagogía.	-0.100	0.497	0.275	0.178
	(0.933)	(0.658)	(0.624)	(0.561)
Porcentaje de profesores con estudios profesionales.	-0.223	0.264	0.103	-0.0104
	(0.645)	(0.459)	(0.457)	(0.418)
Porcentaje de profesores con estudios profesionales en pedagogía.	-0.405	-0.258	-0.276	-0.470
	(0.559)	(0.397)	(0.407)	(0.380)
Porcentaje de profesores con postgrado.	-0.249	-0.230	0.175	0.0506
	(0.931)	(0.672)	(0.710)	(0.655)
Porcentaje de profesores con postgrado pedagógico.	0.652	0.671	1.048*	0.761
	(0.711)	(0.489)	(0.467)	(0.419)
Porcentaje de profesores normalista superiores.	(0.790)	(0.622)	(0.631)	(0.583)
	0.954	0.597	0.877*	0.594
Porcentaje de profesores con estudios técnicos.	(0.711)	(0.489)	(0.467)	(0.419)
Constante	60.79***	8.001***	8.403***	8.091***
	(0.905)	(0.699)	(0.707)	(0.622)
Observaciones	32,643	32,643	32,643	32,643
R-squared	0.246	0.726	0.582	0.576
Robust standard errors in parentheses	*** p<0.01, ** p<0.05, * p<0.1			

Los resultados a nivel de colegio muestran que existe un efecto positivo del índice que se compone de la pensión del colegio, si tiene jornada ordinaria y si es privado (este índice pasa

la prueba de Alpha de Cobranch); este resultado es acorde con la literatura, por un lado varios autores han corroborado que estudiar en colegios privados se traduce en mejores resultados en las pruebas saber 11 (Gabiria y Barrientos, 2001; Ireguri, Melo y Ramos, 2006). Por otro lado es posible que un colegio más costoso se relacione con mejores condiciones físicas, lo cual algunos autores argumentan que repercute positivamente en los resultados educativos (Mediavilla y Calero, 2009). Por ultimo podemos suponer que una mayor exposición al colegio debería generar mejores resultados académicos, por esto es posible que la jornada completa resulte en un mejor puntaje en la prueba saber 11 (García, Maldonado, Perry, Rodrigues y Saavedra, 2014).

Por otra parte las dummys de proporción de educación de los profesores (estructura tomada de García, Maldonado, Perry, Rodrigues y Saavedra, 2014) también muestran que colegios con profesores más calificados tienen un efecto colegio relacionado con mejores resultados, y a pesar de que este efecto no es tan fuerte en la prueba de lenguaje, podemos ver que una institución con mayor proporción de profesores que normalistas tiene un efecto negativo sobre los resultados educacionales.

Un resultado interesante es que el efecto del promedio de la educación de los padres a nivel de colegio es positivo (Ann Owens, 2010; Otto, 1977), lo cual se podría traducir en un efecto par positivo. Por otra parte un curso con más estudiantes en extra edad, o con mayor variación en la edad de los estudiantes tiene un efecto negativo a nivel de colegio; este efecto puede tener dos interpretaciones: 1) la existencia de un modelo a seguir negativo, donde la presencia de estudiantes con atraso académico empeora los resultados de otros estudiantes (Wilson, 1987); 2) podemos estar en presencia de un efecto "focus", donde el profesor al enfrentarse a un curso más homogéneo puede diseñar un programa enfocado a la habilidad media de los estudiantes, generando mejores resultados educativos (Hoxby y Weingarth, 2006). De igual manera es posible que los dos resultados anteriores sufran de problemas de identificación, por lo tanto se deben tomar como un indicio.

Lo más interesante que se puede concluir de estas regresiones es que el efecto del cambio de variaciones de la prueba saber 11 es radicalmente distinto en lenguaje que en matemáticas, se vuelve no significativo en matemáticas, pero en lenguaje si lo es y además es negativo. Esto puede ser evidencia a favor de que el aprendizaje en lenguaje tiene muchos más canales de transmisión sociales, mientras que los canales de trasmisión de matemáticas son más académicos.

Conclusiones:

De nuestro estudio podemos concluir que el barrio tiene efectos sobre la educación de los individuos, esto lo comprobamos con varios métodos de estimación y distintas especificaciones. El elemento a resaltar en este aspecto es que la prueba de lenguaje tiende a ser más sensible a las condiciones contextuales del estudiante que la prueba de matemáticas, que depende principalmente de las condiciones del colegio. Este resultado cobra mayor fuerza si tenemos en cuenta que Bogotá cuenta con una estructura donde barrio y colegio no se superponen exactamente, el vecindario es importante de todas formas. Esta estructura separa mejor el efecto del barrio del efecto del colegio.

Además de lo anterior, podemos ver como el tener en cuenta ambos niveles de anidación permite conocer la significancia real de los efectos fijos, en particular, este efecto es muy claro a la hora de tener en cuenta el género en la prueba de lenguaje. Además, este resultado nos deja rastros de problemas de desigualdad de género en la educación. En general, sobre la primera etapa del modelo, podemos concluir que se mantienen muchos de los resultados encontrados por la literatura, entre otros encontramos que el nivel que explica una mayor porción de la varianza es el de las condiciones familiares y personales, seguido por el colegio, y por último se encuentra el barrio.

Por otra parte podemos ver que la metodología en dos etapas funciona bastante bien, en especial si se tiene una muestra grande pero desbalanceada, pues es muy funcional a la hora de superar los problemas computacionales que traen consigo los modelos jerárquicos multi nivel. Este método permite a su vez, incluir una gran cantidad de variables explicativas para ambos grupos de anidación, lo cual da pie para reforzar la hipótesis de que el lenguaje depende más de condiciones contextuales y nos permite tener indicios sobre otras hipótesis planteadas por la literatura en educación.

Sobre la segunda etapa de nuestro modelo nos muestra efectos muy interesantes, por un lado encontramos evidencia a favor de que la dispersión en resultados de la prueba y en edad de los estudiantes genere efectos negativos a nivel de institución. Por otra parte encontramos evidencia a favor de un efecto espacial, también podemos ver que el crimen afecta negativamente los resultados y que los universitarios en la UPZ afectan positivamente el efecto del barrio.

Bibliografía:

- Mediavilla Bordalejo, Mauro y Calero Martinez, Jorge (2009). "Determinantes internos y externos en el proceso de aprendizaje. Una aproximación al caso español a partir de la ECV-05". *Revista Iberoamericana de Educación*, ISSN: 1681-5653. n.º 50/6 – 25 de octubre de 2009.

- D. Ready, Douglas (2010). "Socioeconomic Disadvantage, School Attendance, and Early Cognitive Development: The Differential Effects of School Exposure". *Sociology of Education*, 2010 DOI: 10.1177/0038040710383520.

- Gaviria Alejandro y Barrientos Jorge Hugo (2001); "Determinantes de la calidad de la educación en Colombia"; Departamento Nacional de Planeación, Dirección de Estudios Económicos; *Documento 159*, 8 de Noviembre de 2001.

- Stephen Gibbons, Olmo Silva y Felix Weinhardt (2013). "EVERYBODY NEEDS GOOD NEIGHBOURS? EVIDENCE FROM STUDENTS' OUTCOMES IN ENGLAND". *The Economic Journal*, © 2013 Royal Economic Society. Accepted Article.

- Goux, Dominique y Maurin, Eric (2003). "Neighborhood Efects on Performances at School." Conference on Changing Condition in Education (Uppsala, 2003).

- Kaztman, Ruben, and Retamoso, Alejandro (2007). "Efectos de la segregación urbana sobre la educación en Montevideo". *Revista De La CEPAL*. (91): 133.

- Owens, Ann. "Neighborhoods and Schools as Competing and Reinforcing Contexts for Educational Attainment". *Sociology of Education* 2010 83: 287, DOI: 10.1177/0038040710383519.

- Small, Mario Luis y Newman, Katherine. "URBAN POVERTY AFTER THE TRULY DISADVANTAGED: The Rediscovery of the Family, the Neighborhood, and Culture." Annu. Rev. Sociol. 2001. 27:23–45.

- Mayer, Susan y Jencks Cristopher. 1989. "Growin up in poor neigborhoods: how much desit matter". *Science. New series*, vol. 243, No 4897.

- Wilson WJ. 1987. "The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy". Chicago: Univ. Chicago Press.

- Brook-Gun et all (2007). "Cognitive and Emotional Outcomes for Children in Poverty" Hanbook of early childhood intervention. capitulo 18.

- Patacchini, Eleonora y Zenou, Yves. 2009. "Neighborhood Effects and Parental involvement in the Intergenerational Transmission of Education". *Journal of Regional Science*, Wiley Blackwell, vol. 51(5), pages 987-1013, December.

- Raundenbush. Stephen W y Bryk, Antony S (2002). "Hierarchical linear models: Applications and Data Analysis Methods." Sage Publications.

- Fielding, Antony (1999). "Why use arbitrary point scores?: Ordered categories in models of educational progress." J. R Statist. Soc. A (1999), 162, part 3. Pp 303 – 328.
- Snijders, Tom A B y Bosker, Roel J (2003). "An introduction to basic and advanced multilevel modeling". Sage Publications.

- Raudenbush, Stephen W (1993), "A crossed random effects model for unbalanced data with application in cross-sectional an longitudinal research". *Journal of Educational Statisticsm*, Winter 1993, Vol. 18, No. 4, pp. 321-349.

- Verbitsky Savitz, Natalya y Raudenbush, Stephen W (2010). *Sociological Methods & Research*. May 2010 38: 515-544.

- Cook, Thomas D., Melissa R. Herman, Meredith Phillips, and Richard A. Settersten, Jr. 2002. ''Some Ways in Which Neighborhoods, Nuclear Families, Friendship Groups, and Schools Jointly Affect Changes in Early Adolescent Development.''Child Development 73(4):1283-309.

- David J. Harding (2009). "Collateral Consequences of Violence in Disadvantaged Neighborhoods" AMERICAN SOCIOLOGICAL REVIEW, VOL. 74 (JUNE: 445–464).

- Rabe-Hesket, Sophia y Skrondal, Andres; Multi level and Longitudinal modeling using Stata. Second Edition. 2008; Stata-press.

- Ireguri, Ana M, Melo, Ligia y Ramos, Jorge. "La educación en Colombia: análisis del marco normativo y de los indicadores sectoriales." Rev. Econ. Ros. Bogotá (Colombia) 9 (2): 175-223, diciembre de 2006.

- Steele, Fiona, Vignoles, Ana Vignoles y Jenkins, Andrew. (2007). The Impact of School Resources on Student Attainment: A Multilevel Simultaneous Equation Modelling Approach. The Journal of Royal Statistical Society, A series, 170(3), 801-824.

- Domínguez Blanco, María Elvira. (2004). Revista electronica de educacion y psicología, Numero 2, Diciembre del 204.

- Fomisano, Michel (2002). "Econometría espacial: características de la violencia homicida en Bogotá". Documentos Cede 2002-10. ISSN 1657-7191. Septiembre.

- Hofmann, David A y Gavin, Mark B. Journal of Management 1998, Vol. 24, No. 5, 623-641.

- Raudenbush, Stephen y Garner, C. L- (1991). “Neighborhood effects on educational attainment: A multilevel analysis.” Sociology of Education, 64 (4) 251-262.

- M. Hoxby, Caroline y Weingarth, Gretchen. (2006. “TAKING RACE OUT OF THE EQUATION: SCHOOL REASSIGNMENT AND THE STRUCTURE OF PEER EFFECTS”. Harvart University Dpt. Of economics. December.

- Aliaga-Linares, Lissette y Álvarez-Rivadulla, María José (2010). Residential Segregation in Bogotá across Time and Scales. Lincoln Institute of Land Policy .Working Paper.

- Rien Vander, Leeden, Busing, Frank y Neijer, Erik. (1997). “Boostrap methods for two level models.” Leiden University. Institute for educational research the Netherlands.

- Garcia Jaramillo, Sandra, Maldonado Carrizosa, Dario, Perry Rubio, Guillermo, Rodriguez Orgales, Katherine y Saavedra Calvo, Juan Esteban. (2014). “Tras la excelencia docente”. Fundación compartir.

- JeremyJ.Albright and Dani M. Marinova; “Estimating Multilevel Models using SPSS, Stata, SAS, and R”; Indiana University 2010.

Anexos:

Anexo 1: Regresiones preliminares

Dado que nuestro objetivo es explicar los resultados en la prueba saber 11 de matemáticas y lenguaje, es importante tener este ordenamiento de los datos en cuenta a la hora de estimar modelos econométricos, pues no tener en cuenta esta estructurar puede llevar a no especificar correctamente la varianza del modelo, subestimándola, lo cual puede llevar a no calcular correctamente la significancia de los efectos fijos. En otras palabras, el modelo no sería eficiente. (Raudenbush y Bryck, 2002).

Para esto comenzamos haciendo estimaciones preliminares usando dos tipos de modelos. Primero usaremos una regresión OLS con tres bases de datos distintas, la primera es una regresión explicando el resultado en matemáticas y lenguaje solo con variables a nivel de familia, esta regresión nos permitirá comparar los coeficientes y la significancia de las variables de la regresión con la base completa, con la base completa sin datos atípicos y nuestra base final. Luego haremos una regresión que tiene en cuenta la anidación de los individuos en colegios, en otras palabras, regresiones que suponen la existencia de efectos fijos o aleatorios por colegios:

En estas regresiones podemos ver que hay un efecto positivo de la educación de los padres, también hay un efecto positivo del ingreso, y en el caso de la prueba de matemáticas existe

regresiones ols

VARIABLES	poblacional matemáticas	poblacional lenguaje	Poblacional sin datos atípicos matemáticas	Poblacional sin datos atípicos lenguaje	Muestral matemáticas	Muestral lenguaje
sexo	2.353***	-0.301***	2.474***	-0.337***	2.658***	-0.224***
	(0.0547)	(0.0444)	(0.0615)	(0.0491)	(0.0948)	(0.0727)
Edad centrada	-0.187***	-0.140***	-0.567***	-0.479***	-0.358***	-0.312***

	(0.00733)	(0.00651)	(0.0274)	(0.0227)	(0.0428)	(0.0331)
Años de educación del padre centrados	0.182***	0.158***	0.178***	0.148***	0.232***	0.190***
	(0.00830)	(0.00687)	(0.00969)	(0.00791)	(0.0165)	(0.0132)
Años de educación de la madre centrados	0.137***	0.141***	0.178***	0.186***	0.209***	0.221***
	(0.00775)	(0.00643)	(0.00934)	(0.00766)	(0.0154)	(0.0123)
Hermano desertor	-1.063***	-1.091***	-0.874***	-1.079***	-0.845***	-1.015***
	(0.119)	(0.103)	(0.137)	(0.118)	(0.215)	(0.178)
Sin hermano	-0.116*	0.0294	-0.145**	0.0772	-0.156	0.0558
	(0.0595)	(0.0484)	(0.0663)	(0.0532)	(0.102)	(0.0782)
Ingreso familiar centrado	1.432***	1.146***	1.466***	1.148***	1.491***	1.083***
	(0.0276)	(0.0209)	(0.0301)	(0.0227)	(0.0454)	(0.0329)
Controles laborales	si	Si	si	si	si	si
Constante	45.01***	47.28***	45.43***	47.57***	46.27***	48.35***
	(0.0828)	(0.0714)	(0.104)	(0.0894)	(0.164)	(0.137)
Observaciones	95,764	95,764	76,536	76,536	32,692	32,692
R-squared	0.159	0.149	0.163	0.160	0.183	0.186

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

un efecto positivo asociado al género masculino. Por otra parte estudiantes en extra edad o con hermanos desertores tienen efectos negativos sobre ambas pruebas.

El paso a seguir es nuestro modelo con efectos de colegio, para esto primero debemos saber si nuestra estructura de datos debe ser modelada con efectos fijos o con efectos aleatorios, pues nuestro modelo ideal debe tener en cuenta una doble anidación y para esto usamos modelos jerárquicos lineales, que suponen la utilización de efectos aleatorios. Esto lo podemos saber usando un test de "Hausman de endogeneidad" (Durbin-Wu-Hausman) este test compara dos modelos de panel o con estructura anidada (para este primer caso anidados en colegios), uno con efectos fijos y otro con efectos aleatorios estimado por GLS, si se

acepta la hipótesis nula quiere decir que el modelo que mejor explica la variable dependiente es el modelo con efectos fijos, en nuestro caso:

	Hausman	**Colegios**		
	poblacional		submuestra	
	lenguaje	matemáticas	lenguaje	matemáticas
chi2	1597.49	1789.24	783.75	714.58
prob	0.00	0.00	0.00	0.00

Lo anterior indica que nuestros datos deben ser modelados con efectos aleatorios, la gran debilidad del test anterior es que supone que el modelo es verdadero y prueba esto de manera asintótica, en otras palabras, supone que el número de observaciones tienden a infinito, pero que la forma funcional es la correcta, dado esto lo ideal sería ir haciendo esta prueba según las distintas especificaciones. A continuación mostraremos nuestras regresiones con efectos aleatorios, para estas regresiones usamos máxima verosimilitud con el objetivo de hacer más comparable la estimación del efecto entre grupos, pues este es el método de estimación de los modelos multi nivel:

Regresiones panel: colegios	**máxima verosimilitud**			
	poblacional sin datos atípicos	poblacional sin datos atípicos	Muestral	Muestral
VARIABLES	matemática	lenguaje	matemática	lenguaje
sexo	2.673***	-0.147***	2.804***	-0.0341
	(0.0614)	(0.0507)	(0.0949)	(0.0757)
Edad centrada	0.780***	-0.606***	-0.715***	-0.518***
	(0.0265)	(0.0219)	(0.0417)	(0.0333)
Años de educación del padre centrados	0.0803***	0.108***	0.119***	0.103***
	(0.00901)	(0.00745)	(0.0163)	(0.0131)
Años de educación de la madre centrados	0.0902***	0.0756***	0.110***	0.145***
	(0.00921)	(0.00762)	(0.0153)	(0.0123)

Hermano desertor	-0.635***	-0.854***	-0.613***	-0.772***
	(0.136)	(0.112)	(0.224)	(0.179)
Sin hermano	-0.229***	-0.0482	-0.201**	-0.0318
	(0.0621)	(0.0513)	(0.0939)	(0.0751)
Ingreso familiar centrado	0.485***	0.441***	0.595***	0.507***
	(0.0308)	(0.0256)	(0.0449)	(0.0361)
Controles laborales	si	Si	si	si
Constant	45.82***	48.00***	46.03***	48.24***
	(0.158)	(0.121)	(0.212)	(0.161)
Observations	76,536	76,536	32,692	32,692
Numero de colegios	1,437	1,437	1,181	1,181
Standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

De igual manera la mayor ventaja de esta aproximación es ver la cantidad de varianza explicada al interior de los grupos o entre grupos, tabla que mostramos a continuación:

	efectos poblacional sin datos atípicos	Colegios		
	lenguaje	matemáticas	submuestra lenguaje	matemáticas
Within	6.388682	7.724629	6.11121	7.637153
between	2.938119	4.193379	2.697126	4.011032

varianza[2] explicada por los grupos	.1745789	.2276174	.1630269	.2161995

Como podemos ver la mayor parte de la varianza en todos los casos se encuentra al interior de las mismas instituciones educativas. Este resultado se mantiene con el cambio en la muestra. Esto implica que el nivel que mayor parte de la varianza explica es el nivel familiar y personal.

De todas maneras, es muy posible que parte de la varianza de la variable dependiente se explique por la pertenencia a un barrio determinado. No solo las interacciones académicas construyen el conocimiento, incluso en el grupo etario que estamos analizando, pueden existir ciertas interacciones sociales fuera del aula de clases que tengan un efecto sobre los resultados académicos (Kaztman y Retamoso 2007; Patacchini y Zenou 2009; Goux y Maurin 2003; Owens 2010). Por esto es importante analizar la estructura de los efectos con los que debemos analizar la anidación en barrios (sean fijos o aleatorios):

	Hausman submuestra: lenguaje	**Barrios** matemáticas
chi2	915.59	1052.22
prob	0	0

Al igual que en el caso anterior nuestro modelo de efectos barrio debe ser modelado con efectos aleatorios, para esta prueba el estimador por GLS calcula que la varianza explicada por los barrios es igual a cero al tener en cuenta variables como el ingreso y las variables dicotómicas de ocupación, por esto la realizamos con y sin estas variables. De igual manera para ambos modelos (el que tiene variables de ingreso y ocupación y el que no) la prueba de Hausman encuentra que la diferencia entre coeficientes es no sistemática, en otras palabras que es apropiado usar efectos aleatorios; además de lo anterior hay que tener en cuenta dos motivos que nos llevan a pensar que se debe modelar el efecto vecindario con efectos aleatorios: 1) el estimador por cuasi máxima verosimilitud, que es el que más nos interesa, si encuentra que los barrios explican la varianza de la variable dependiente; y 2) el hecho de que al tener en cuenta los controles de ingreso y ocupación el efecto fijo sea nulo puede ser indicio de un sesgo de selección de barrios, que podría estar estrechamente relacionado al

[2] rho = (between)^2/[(between)^2 + (witin)^2]

sesgo de selección de colegios; en otras palabras, al controlar por el ingreso y la ocupación delos padres, se puede estar controlando por la decisión de esas familia sobre elegir un barrio y un colegio determinado. Lo anterior refuerza la idea de usar el modelo con doble anidación, pues este tiene en cuenta la celda (grupos creados por los cruces de barrio y colegio) en la cual estaría reflejada esta decisión de elegir un barrio y un colegio dado. Ahora nuestras regresiones con efecto vecindario y efectos aleatorios usando máxima verosimilitud:

regresiones panel barrios	**Muestral**	
VARIABLES	lenguaje	matemática
sexo	-0.222***	2.660***
	(0.0718)	(0.0931)
Edad centrada	-0.462***	-0.585***
	(0.0329)	(0.0427)
Años de educación del padre centrados	0.154***	0.194***
	(0.0134)	(0.0173)
Años de educación de la madre centrados	0.184***	0.170***
	(0.0126)	(0.0163)
Hermano desertor	-0.944***	-0.808***
	(0.185)	(0.240)
Sin hermano	0.0292	-0.189*
	(0.0772)	(0.100)
Ingreso familiar centrado	0.780***	1.077***
	(0.0339)	(0.0438)
Controles laborales	si	si
Constant	48.67***	46.59***
	(0.203)	(0.277)
Observations	32,692	32,692

Number of UPZ	110	110

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Como podemos ver se repiten los resultados de las estimaciones anteriores, dado que los estimadores de las variables a primer nivel no son solo una de nuestras hipótesis a poner a prueba, sino que además pueden ser, una prueba en si misma de nuestra especificación para nuestro modelo de efectos cruzados; es importante asegurarnos que la especificación de estos estimadores sea la correcta. A continuación mostramos los efectos al interior de los barrios y entre estos estimados por el modelo anterior:

	efectos de barrio Submuestra re	
	lenguaje	matemáticas
witin	6.430191	8.34033
between	1.448449	2.072314
varianza explicada por los grupos	.0482907	.058147

Lo interesante del ejercicio anterior es que encontramos que la varianza explicada por la anidación en barrios es menor que la varianza explicada por la anidación en colegio, usando el método de cuasi máxima verosimilitud, de igual manera este resultado es de esperarse, incluso Gibbons, Olmo y Weinhardt (2012) encuentra que cambios en la composición del barrio no tiene ningún efecto, por otra parte Katsman y Ratmoso (2007) usando una descomposición de varianza similar a la de este estudio encuentran que el nivel barrio explica más varianza que el nivel colegio; como se puede ver aún no existe consenso sobre el efecto vecindario, pero la mayoría de estudios encuentran que las condiciones del colegios tienen una mayor correlación con los resultados educativos que las condiciones del barrio. Es importante utilizar este método de estimación, pues nuestras especificaciones multi nivel utilizan este mismo método lo cual hace que nuestros ejercicios sean más comparables.

De los ejercicios de regresión anteriores podemos decir que en ambos casos, tanto en el panel de colegios como en el de barrios, se explica más la varianza al interior de los grupos que

entre grupos, Esto implica que el nivel de agregación que más varianza explica es el nivel de condiciones familiares y personales.

Anexo 2: Descripción y comparación de los controles familiares, base sin datos atípicos y la base poblacional.

		Sub muestra geo referenciada		Base poblacional	85083 observaciones
Variable	Mean	Std. Dev.	Mean	Std. Dev.	
matematica	47.70142	9.357431	46.77588	9.130758	
lenguaje	48.29043	7.188256	47.53419	7.332603	
sexo	0.4704714	0.4991345	0.4702937	0.4991197	
edan centrada	-0.01885	1.140865	-2.42E-09	1.179862	
m_priminc	0.0831463	0.2761071	0.0953422	0.2936887	
h_priminc	0.1088471	0.3114517	0.1173795	0.3218738	
m_primaria	0.1168192	0.3212093	0.1319065	0.3383911	
h_primaria	0.1328786	0.3394484	0.1412386	0.3482695	
m_secundaria	0.2530075	0.4347415	0.2431391	0.4289809	
h_secundaria	0.2349335	0.4239634	0.2246865	0.4173781	
m_tecinc	0.0276291	0.16391	0.0262097	0.1597594	
h_tecinc	0.0186496	0.1352862	0.0163605	0.1268583	
m_tecnico	0.0793185	0.2702391	0.0701668	0.2554294	
h_tecnico	0.0605537	0.2385133	0.0528425	0.2237203	
m_profinc	0.0401773	0.1963777	0.0348248	0.183337	
h_profinc	0.034709	0.1830445	0.0298414	0.1701507	
m_profecional	0.1558453	0.3627138	0.1272287	0.333231	
h_profecional	0.1691706	0.3749079	0.1370897	0.343944	
m_postgr	0.0491855	0.2162583	0.0383155	0.191958	
h_postgr	0.0589133	0.2354657	0.0434752	0.2039255	
hermano_desertor	0.0418178	0.2001754	0.0494693	0.2168471	
no_hermano	0.3363265	0.4724589	0.3227319	0.4675239	

cent_ing_fam	0.0105104	1.531444	-4.77E-09	1.439862
m_otraocup	0.0521211	0.2222744	0.0578729	0.2335043
h_otraocup	0.0549704	0.2279257	0.0574615	0.2327238
m_estudian	0.0032522	0.0569358	0.0031851	0.0563473
h_estudian	0.0007771	0.0278655	0.0007405	0.0272014
m_renta	0.0042595	0.0651265	0.0041254	0.0640969
h_renta	0.0044897	0.0668558	0.0039491	0.0627179
m_jubilados	0.0161457	0.1260378	0.0137983	0.1166536
h_jubilados	0.04271	0.2022053	0.0376338	0.1903102
m_hogar	0.3757267	0.4843169	0.3739172	0.4838449
h_hogar	0.0069361	0.0829948	0.0070637	0.0837489
h_cuentapropia	0.2105854	0.4077303	0.2116286	0.4084653
m_cuentapropia	0.1145743	0.3185121	0.1201415	0.3251289
m_indep_prof	0.040868	0.1979872	0.0368229	0.1883278
h_indep_prof	0.1028895	0.3038189	0.0924509	0.2896631
m_auxiliar	0.0697059	0.2546543	0.0699905	0.2551325
h_auxiiar	0.0556899	0.2293251	0.0543234	0.2266561
m_emple_tecprof	0.0744258	0.2624664	0.0654067	0.2472436
h_emple_tecprof	0.0794912	0.2705077	0.0704253	0.255864
m_direc	0.0393139	0.1943435	0.033579	0.1801439
h_direc	0.0563806	0.2306585	0.0494458	0.2167983
m_gernetes	0.0209808	0.1433221	0.0186641	0.1353366
h_gerentes	0.0367812	0.1882269	0.0327562	0.177999
m_peqempre	0.0602084	0.2378759	0.0568621	0.2315803
h_peqempre	0.0835492	0.2767145	0.0769601	0.2665298
m_empresarios	0.0213838	0.144662	0.0190637	0.1367499
h_empresarios	0.0412709	0.198919	0.0348718	0.1834562
educación centrada padre	0.6446561	2.821774	2.49E-08	3.818524
educación centrada padre	0.401237	2.533227	-9.07E-11	3.182365

Anexo 3: Variables a nivel de colegio.

Variable	Obs	Mean	Std. Dev.
cv_leng	34829	0.1327969	0.0301382
educ_c~madre	34829	0.3992061	1.691607
educ_c~padre	34829	0.6427056	1.860677
dummy_pent~2	34829	0.166614	0.3726363
dummy_pent~3	34829	0.0837808	0.2770627
dummy_pent~4	34829	0.0691378	0.2536919
dummy_pent~5	34829	0.0886617	0.284259
dummy_pent~6	34829	0.1350312	0.3417617
ordinaria	34829	0.2950415	0.456068
calen_a	34829	0.9461943	0.2256372
privados	34829	0.4856872	0.4998023
postgrado_~e	34829	0.1891768	0.1398291
profe~l_cole	34829	0.1207353	0.0912717
profe~g_cole	34829	0.204378	0.098446
norm_tec_c~e	34829	0.1844207	0.1497653
cole_inst_~n	34829	5.37397	5.06528

Anexo 4: Variables a nivel de barrio.

Variable	Obs	Mean	Std. Dev.	Min	Max
univer_upz	34746	39.40707	43.38125		
educ_upz	34746	7.819763	1.792066		
desert_upz	34746	0.1435643	0.0599791		

cv_educ_secc	34743	0.128035	0.0583343
despvio_upz	34746	0.006657	0.0028457
labusco_upz	34746	8.535791	2.097054
clust_lh_l~g	34829	0.0015217	0.0389801
clust_hl_l~g	34829	0.0003445	0.0185589
clust_hh_l~g	34829	0.1015533	0.3020643
clust_ll_l~g	34829	0.0129777	0.1131797
clust_lh_mat	34829	0.0015217	0.0389801
clust_hl_mat	34829	0.0003445	0.0185589
clust_hh_mat	34829	0.0919349	0.2889382
clust_ll_mat	34829	0.0850154	0.2789086

.

Anexo 4.1: Local Moran´s Index.

El objetivo central de este índice es encontrar espacialmente (en nuestro caso para la ciudad de Bogotá) concentración de clúster de altos y bajos resultados en la prueba saber 11, y a su vez encontrar la presencia de grupos atípicos, que son básicamente barrios con altos resultados rodeados de barrios con bajos resultados y viceversa, la ecuación es la siguiente:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{j,i}(x_i - \bar{X})$$

Donde x_i sería el resultado de la prueba saber 11 de la UPZ i, $w_{j,i}$ es la distancia entre la unidad j y la unidad i, y $\bar{X}$ es el promedio de la prueba saber 11. Para la ecuación anterior S_i^2 es:

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n}(x_i - \bar{X})}{n-1} - \bar{X}$$

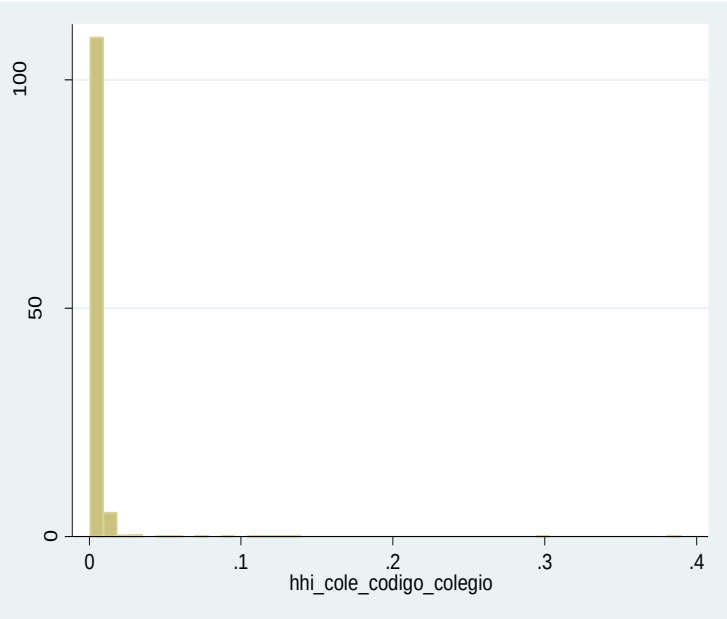
Anexo 5:

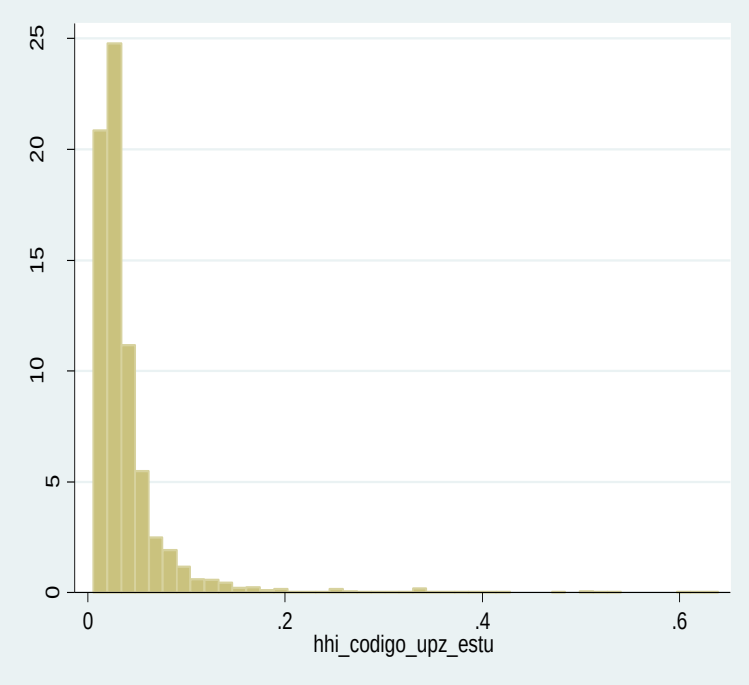
Dado que los datos de nuestra muestra son anidados, pero no tienen una anidación perfecta, sino que poseen una anidación cruzada, la mejor manera de estimar este tipo de anidación en los datos es usar un modelo jerárquico de efectos aleatorios cruzados. Desafortunadamente estos modelos demandan gran poder de computación y dado que tenemos una base de datos con muchas observaciones, es posible que los algoritmos de estimación no converjan. Los algoritmos más usados son el EM y el PQL2; el problema con estos algoritmos está en la integración necesaria para encontrar la distribución latente de los datos. Para esta integración se podría usar una aproximación de Laplace pero se necesita suponer puntos de integración, los cuales se puede hallar por medio de simulaciones, o en caso que la variable dependiente sea discreta se pueden usar puntos de integración arbitrarios (Fielding, 1999)[3].

A pesar de lo anterior, nuestra base de datos es tan grande que muy probablemente ninguno de esto métodos funcione, en especial si tenemos en cuenta que por ser una anidación cruzada con muchos barrios y colegios que contienen estudiantes, el estimar el efecto aleatorio de los cruces colegio-barrio específicos (de cada celda) es muy demandante computacionalmente. Por esta razón solo estimaremos modelos jerárquicos lineales de efectos cruzados sin variables explicativas (lo cual reduce el número de efectos aleatorios a estimar), o solo con variables explicativas al primer nivel (nivel familiar), lo cual no debería presentar ningún problema de convergencia. Es también por esto que el número de grupos, o barrios no debe ser tan grande, pues a mayor número de barrios, mayor número de celdas.

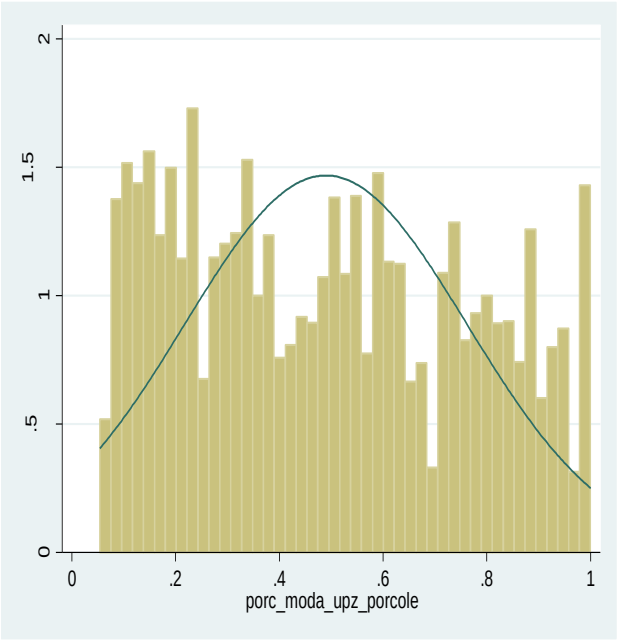
Anexo 6: Histograma del índice de Herfindal, concentración de colegios por UPZ y concentración de barrios por colegios:

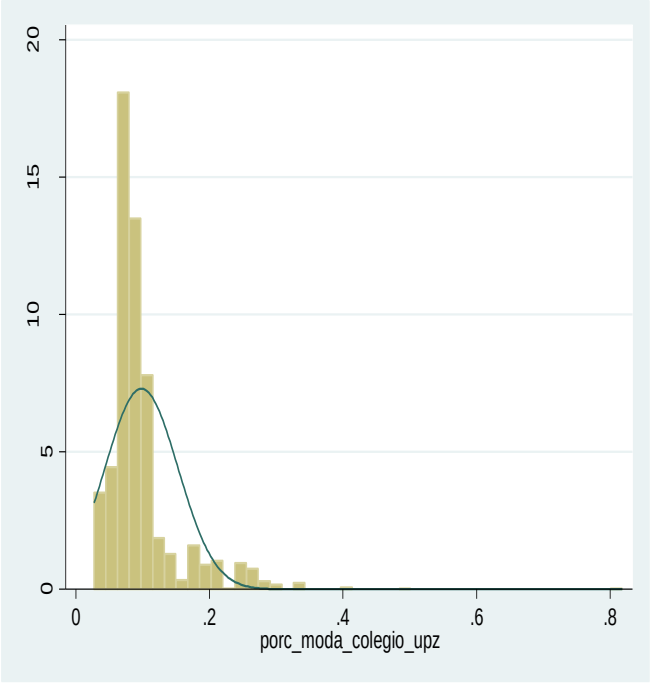
[3] Este proceso también se puede entender por medio de econometría bayesiana, y en efecto el cálculo preciso de los efectos aleatorios se hace por métodos bayesianos.





Histograma de la UPZ moda por colegios y Histograma del colegio moda por UPZ





Como podemos ver, rara vez un barrio asiste siempre al mismo colegio y viceversa, esto es sumamente positivo para nuestro modelo pues implica que ninguna de las estimaciones de grupo esta sesgada por correlacionar perfectamente uno de los niveles de anidación con el otro.

Anexo: tabla de variables.

VARIABLES	Descripción:
Familia:	
sexo	Variable Dummy que indica 1 si el estudiante es hombre y 0 si es mujer
Edad centrada del estudiante	Edad en años del estudiante, centrada con la media general de la muestra o población.
Años de educación de la madre centrados	Educación de la madre en años centrada respecto a la media general de la muestra o población.
Años de educación del padre centrados	Educación de padre en años centrada respecto a la media general de la muestra o población.
Hermano desertor menor de 15 años	Hermano menor de 15 años que abandonó sus estudios.
Sin hermanos	Estudiante sin hermanos.
Ingreso familiar centrado	Ingreso familiar auto reportado, su escala es la original del cuestionario, va de 1 a 7 donde uno es menos de un salario mini y 7 es más de 10 salarios mínimos. Centrado respecto a la muestra o población.
m_otraocup	Estudiantes que reportaron que su madre tiene otra ocupación.
h_otraocup	Estudiantes que reportaron que su padre tiene otra ocupación.
m_estudian	Estudiantes que reportaron que su madre se encuentra estudiando.

h_estudian	Estudiantes que reportaron que su padre tiene otra ocupación.
m_renta	Estudiantes que reportaron que su madre vive de la renta.
h_renta	Estudiantes que reportaron que su padre vive de la renta.
m_jubilados	Estudiantes que reportaron que su madre esta jubilada.
h_jubilados	Estudiantes que reportaron que su padre está jubilado.
m_hogar	Estudiantes que reportaron que su madre vive del hogar.
h_hogar	Estudiantes que reportaron que su padre vive del hogar.
h_cuentapropia	Estudiantes que reportaron que su padre es cuenta propia.
m_cuentapropia	Estudiantes que reportaron que su padre es cuenta propia.
m_indep_prof	Estudiantes que reportaron que su madre trabaja como independiente profesional.
h_indep_prof	Estudiantes que reportaron que su padre trabaja como independiente profesional.
m_auxiliar	Estudiantes que reportaron que su madre es auxiliar.
h_auxiiar	Estudiantes que reportaron que su padre es auxiliar.
m_emple_tecprof	Estudiantes que reportaron que su madre trabaja como empleado técnico o profesional.
h_emple_tecprof	Estudiantes que reportaron que su padre trabaja como empleado técnico o profesional.
m_direc	Estudiantes que reportaron que su madre tiene nivel de en una empresa

h_direc	Estudiantes que reportaron que su padre tiene nivel de en una empresa
m_gernetes	Estudiantes que reportaron que su madre tiene nivel de gerente.
h_gerentes	Estudiantes que reportaron que su padre tiene nivel de gerente.
m_peqempre	Estudiantes que reportaron que su madre es dueño de una pequeña empresa.
h_peqempre	Estudiantes que reportaron que su padre es dueño de una pequeña empresa.
m_empresarios	Estudiantes que reportaron que su madre es dueño de una empresa.
h_empresarios	Estudiantes que reportaron que su padre es dueño de una empresa.
alt_calif_mad	Dummy que indica 1 si reporto que la madre es empresaria, gerente, directivo, profesional independiente o empleada técnica o profesional.
alt_calif_pad	Dummy que indica 1 si reporto que el padre es empresario, gerente, directivo, profesional independiente o empleada técnico o profesional.

Barrio:	Los datos originalmente estaba a nivel de barrio y se promediaron para agregarlos a nivel de UPZ.
educ_upz	Promedio de años de educación en la UPZ.
univer_upz	Promedio de universitarios en la UPZ.
count_hom, clust_lh_mat	Promedio de homicidios por UPZ.
clust_hh_leng, clust_lh_mat	Clúster de altos resultados en matemáticas y en lenguaje a nivel de UPZ usando el "local morans index"
clust_ll_leng, clust_lh_mat	Clúster de bajos resultados en matemáticas y en lenguaje a nivel de UPZ usando el "local morans index"

clust_hl_leng, clust_lh_mat	UPZ de altos resultados en matemáticas y en lenguaje, rodeada de UPZ con bajos resultados usando el "local morans index"
clust_lh_leng, clust_lh_mat	UPZ de bajos resultados en matemáticas y en lenguaje, rodeada de UPZ con altos resultados usando el "local morans index"

Colegio:	
cv_edad	Cambio de variaciones de edad al interior de la cede.
cv_leng	Cambio de variaciones de la prueba de lenguaje al interior de la cede.
cv_mat	Cambio de variaciones de la prueba de matemáticas al interior de la cede.
zprivados_pension_ord	Índice que incluye si los colegios tienen jornada ordinaria, si son privados y el valor de la pensión estandarizada, este índice paso la prueba de Alpha de Cobranch
educ_cole_años	Años de educación de los padres sumados y promediados por el número de estudiantes de la cede
BACHILLERATO_PEDAGOGICO	Variable que indica el porcentaje de profesores con bachillerato pedagógico a nivel de sede.
OTRO_BACHILLERATO	Variable que indica el porcentaje de profesores con estudios técnicos en pedagogía a nivel de sede.
TECNICO_PEDAGOGICO	Variable que indica el porcentaje de profesores con estudios técnicos en pedagogía a nivel de sede.
TECNICO_OTRO	Variable que indica el porcentaje de profesores con estudios técnicos a nivel de sede.
PROFESIONAL_PEDAGOGICO	Variable que indica el porcentaje de profesores con título de profesional en pedagogía a nivel de sede.

PROFESIONAL_OTRO	Variable que indica el porcentaje de profesores con título de profesional a nivel de sede.
POSTGRADO_PEDAGOGICO	Variable que indica el porcentaje de profesores con título de postgrado en pedagogía a nivel de sede.
POSTGRADO_OTRO	Variable que indica el porcentaje de profesores con título de postgrado a nivel de sede.
tema_matematica_2006	Rezago del resultado en matemáticas a nivel de sede.
tema_lenguaje_2006	Rezago del resultado en lenguaje a nivel de sede.

Sobre UPZ:

El primer elemento a tener en cuenta es el tamaño de los grupos, esto implica no solo tener en cuenta el tamaño de los barrios y de los colegios, sino también de las combinaciones de barrios y colegios y los grupos que estas combinaciones generan. En otras palabras, dado que no todos los estudiantes de un colegio viven en la misma UPZ y viceversa, se generan múltiples grupos de estudiantes que asisten a la misma escuela y viven en el mismo barrio. Para ser más específicos, en nuestros datos existen 9571 cruces entre barrios y colegios (que de ahora en adelante denominaremos celdas), las cuales tienen en promedio 16 estudiantes, de igual manera la desviación típica es de casi 19, lo cual implica que nuestra muestra no es nada balanceada.