CONFORMACIÓN DE LA BASE DE DATOS PANEL DE LA PRUEBA ENLACE Y PLANEA Y CHEQUEO DE CALIDAD DE LA INFORMACIÓN

REPORTE PARA EL BANCO MUNDIAL

Bernardo García Bulle

8 de julio de 2019

Objetivos

Este reporte tiene los siguientes objetivos:

1. Formar una base de datos panel que siga el desempeño de los alumnos en la Prueba Enlace a través de los años. Este ejercicio involucra usar CURP para seguir a los alumnos en el tiempo. El reto principal es saber cómo manejar adecuadamente los CURP cuando tienen errores o están incompletos. En este caso, si se quisiera aumentar la proporción de *matches* habría que estar dispuesto a incurrir en el riesgo de errores de *matching*. Para dejar que el usuario tome esta decisión creamos una base de datos que solo usa *matches* exactos y otra en donde adicionalmente incluimos *matches* imperfectos (*fuzzy matching*): es relativamente poco lo que ganamos con *fuzzy matches*.

2. Evaluar la calidad del panel resultante mediante la siguiente forma: Primero, evaluamos si existen las observaciones en el tiempo, es decir, si es posible seguir a los estudiantes a través del tiempo y si existen lagunas en los datos. Segundo, mostramos las distribuciones de calificaciones, que deberían ser suaves. Tercero, exploramos si en una misma escuela parece haber cambios bruscos en el tamaño de su matrícula en el tiempo, es decir, en el número de estudiantes que tiene (esto debería ser muy atípico). Cuarto, observamos si existen cambios atípicos en las calificaciones promedio a nivel escuela de un año a otro, y en la persistencia a nivel alumno de las calificaciones. Quinto, exploramos los cierres y aperturas de escuelas, y también sobre la distancia que hay entre escuelas a las que dejan de ir los alumnos porque cierran, y las escuelas hacia donde se cambian

3. Adicionalmente, creamos una base panel con datos de la encuesta de primarias de inicio de cursos del 911 para los años 2006 a 2012. Se muestran datos descriptivos sobre el panel creado y, además, se hizo un análisis comparativo utilizando el panel exacto Enlace y el panel 911 para derivar la siguiente información: proporción relativa de estudiantes que presentaron la prueba Enlace entre el panel exacto Enlace y el panel 911; proporción relativa de estudiantes atrasados que presentaron la prueba Enlace entre el panel exacto Enlace y el panel 911, y proporción relativa de estudiantes repetidores que presentaron la prueba Enlace entre el panel exacto Enlace y el panel 911.

Entregables

Objetivo 1: Bases de datos Enlace anonimizadas

1. Base de datos panel Enlace con match de CURP exactos: panel_exacto_a.dta
2. Base de datos panel Enlace con match de CURP exactos agregando las observaciones para 3ero de preparatoria: panel_exacto_18_con_prepa_a.dta
3. Base de datos panel Enlace con match de CURP *fuzzy*: panel_fuzzy_a.dta
4. Do file de Stata que crea las tres bases de datos mencionadas: genera_panel_ENLACE.do

Objetivo 2: Calidad del panel y los datos

1. Reporte del análisis de la calidad del panel exacto y los datos
2. Do file que hace el análisis y gráficas del reporte: reporte.do

Objetivo 3: Análisis comparativo usando panel exacto Enlace y panel 911

1. Base de datos panel 911: panel911_primaria_0616.dta
2. Do file que crea el panel 911: limpiaf911.do
3. Do file que hace el análisis de comparación entre la base de datos panel Enlace y 911: reportef911.do

Objetivo 1: Generación de las bases de datos panel

Partimos de 18 bases de datos entregadas por el Banco Mundial como muestra la tabla 1.

La base de Enlace del 2011 estaba incompleta y la de 2010 tenía solamente cerca de la mitad de las calificaciones. Conseguimos los faltantes por nuestra parte y así pudimos completar los datos. Una primera observación es que el número de folios es mayor al número de CURP, esto se da porque **cerca de 5 %-10 % de los CURP no existen.** En el año 2006 la situación es más crítica pues faltan cerca del 23 % de los CURP.

Tabla 1: Bases de datos originales

Nombre	Folios	CURP de 18 dígitos	CURP de 16 dígitos	CCT	variables	Tamaño en KB
ENLACE2006 (1)	9529490	-	-	111316	15	969118
enl06nal_nombres	9218490	7460501	8377058	-	6	999271
enl07_A	3966280	-	-	45876	20	547981
enl07_B	6182386	-	-	74020	20	857540
enl07nal_nombres	10148666	9507138	10860888	-	6	1139745
RESULT_ALUMNOS_08_A	4306540	-	-	51539	21	407949
enl08_B	5646800	-	-	68433	23	842895
enl08nal_nombres	9910885	9719469	10130905	-	7	909789
RESULT_ALUMNOS_09_A	8029920	-	-	88285	30	846912
RESULT_ALUMNOS_09_B	5157768	-	-	29496	32	946774
enl09nal_nombres	13187682	12735721	13315650	-	8	1455283
RESULT_ALUMNOS_10 _A	6054266	-	-	52526	8	266059
RESULT_ALUMNOS_10 _B	6054266	-	-	67379	8	278878
RES_ENLACE_10_2.csv[1]	13772359	-	-	119905	30	2495186
enl10nal_nombres	-	13537621	14004223	-	3	1156664
resul_enlace_11	8759180	-	-	90538	33	1152000
alumnos_curp_11	8758989	8638956	8986664	-	3	480000
resul_alum_eb12	13507167	-	-	114346	32	1411401
nombres_enlb_12_nac	13507167	13307167	13711727	-	8	1925829

Nota: Por cada año, existen de una a dos bases con resultados identificados por un folio y una base que mapea los folios con los CURP. Los últimos dos dígitos del CURP son los identificadores que en muchas ocasiones no aparecen o son reemplazados por dos asteriscos. El CCT es el identificador único de las escuelas.

La Tabla 2 nos muestra los datos que tenemos y el panorama general para el seguimiento de las generaciones que presentaron la prueba Enlace y que se encuentran capturadas en las bases de datos que nos otorgaron. En esta tabla se denota con las casillas en blanco la ausencia de base de datos para ese año y grado en particular.

Nótese que sólo se tienen datos de 3er grado de preparatoria; esto se debe a que la prueba Enlace únicamente se aplicó a dicho grado. Además, para 2011 no contamos con información de secundarias y para los años 2006, 2007 y 2008 no hay datos de primero y segundo de secundaria.

Asimismo, cada generación tiene un color y número correspondiente; esto con el propósito de que el lector pueda identificar el seguimiento retrospectivo que se llevó a cabo para una generación en específico.

Tabla 2: Generaciones y años calendario

grado	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006
3ro prim					13	12	11	10	9	8	7
4to prim					12	11	10	9	8	7	6
5to prim					11	10	9	8	7	6	5
6to prim					10	9	8	7	6	5	4
1ro secu					9		7	6			
2do secu					8		6	5			
3ro secu					7		5	4	3	2	1
1ro prepa											
2do prepa											
3ro prepa	8**	7**	6	5	4	3	2	1	0		

Nota: Las generaciones se pueden seguir por color y por los números dentro de las celdas.

** a partir de 2014 se implementó la prueba PLANEA que no fue censal por lo que el seguimiento de muchos alumnos se ve interrumpido. Además los resultados de la prueba PLANEA no son comparables con los resultados de ENLACE porque las pruebas fueron diseñadas de distinta manera.

Por ejemplo, la generación 6 puede seguirse desde cuarto de primaria hasta segundo de secundaria, y, finalmente, saltando hasta 3ero de preparatoria; mientras que la generación 12 solo puede seguirse de tercero a cuarto de primaria. Para las generaciones que llamamos 0 y 13 solo observamos un año.

En el panel **panel_exacto_18_con_prepa_a.dta** contamos con estas 14 generaciones en caso de que se quieran usar los datos de preparatoria como en los años 2015 y 2016 se presentó la prueba PLANEA que no es censal ni es comparable con la prueba ENLACE, la base de datos cubre únicamente de 2006 a 2014 para todos los grados que tuvieran información disponible, de

3ro de primaria a 3ro de preparatoria. Para el reporte nos centramos en los datos de secundaria y primaria.

Aunque se puede dar seguimiento a generaciones, no necesariamente pueden seguirse todos los estudiantes. Hay varios casos que si se dan interrumpen el seguimiento:

1. El estudiante deserta.
2. El estudiante no asiste a la prueba Enlace.
3. CURP deficientes que impiden el seguimiento del estudiante.

Es difícil distinguir estos casos, aunque algo puede aprenderse sobre el caso 1 versus el caso 2 si vemos que en un año el estudiante deja de ir, pero al siguiente "vuelve". Para intentar resolver el caso 3 se utilizo el fuzzy match utilizando únicamente los primeros 16 dígitos del CURP.

El do file adjunto a este entregable llamado **genera panel ENLACE.do** toma estas 18 bases y genera las siguientes dos bases de datos panel que se describen en la Tabla 3. Puesto que para muchos años (2013 a 2016) únicamente observamos el tercer grado de preparatoria, las siguientes bases de datos se componen únicamente de alumnos que asistían a la **primaria o secundaria**.

Tabla 3: Paneles de datos armados por nosotros

Nombre	CURP	CCT	obs.	variables	prim. com.	sec. com.	Años prom.
panel_fuzzy	30,435,074	146,068	85304,472	11	5476,705	32,315	2.8
panel_exacto	30,805,408	131,257	74510,112	11	4602,253	26,489	2.4

Nota: Nombre se refiere al nombre de la base de datos, CURP al número de CURP únicos en la base, CCT el número de CCT únicos en la base, obs. el número de observaciones que tiene la base de datos. Prim. com. se refiere al número de alumnos que presentan la prueba durante los 4 años de primaria, sec. com. se refiere al número de alumnos que presentan la prueba durante los 3 años de secundaria y años prom. el promedio de veces que todos los alumnos de la base presentan la prueba. La principal diferencia entre panel_exacto y panel_fuzzy es que el segundo solo considera los primeros 16 caracteres del CURP para hacer el match

Para hacer el Panel_ fuzzy nos quedamos con los primeros 16 caracteres del CURP. Con este nuevo CURP buscamos matches adicionales, es decir, primero se hace un match exacto y después un fuzzy match. Al usar solo 16 dígitos podríamos estar matcheando el CURP de dos personas distintas de manera errónea, aun así, son pocos los matches que se recuperan. Pasamos de 2.4 años en promedio por alumno a 2.8 y se hacen únicamente alrededor de 400,000 matches más que se pueden ver en la columna CURP de la tabla 3.

Para hacer el panel_exacto, a diferencia del panel_fuzzy, nos quedamos únicamente con los CURP que cumplían con las especificaciones estándar (donde el orden de los dígitos y el hecho de si el dígito es número o letra son de suma importancia) y cuyo número de dígitos fuera 18.

Tabla 4: Años que deberíamos tener de cada generación si no hubieran faltas ni deserciones y años promedio que efectivamente encontramos con panel exacto

Generación	Años observados	Media de años observados	% Observados	Número de niños
1	1	1.00	100 %	1,398,084
3	1	1.00	100 %	1,585,198
4	2	1.68	84 %	1,272,749
5	4	3.09	77 %	1,231,146
6	5	3.91	78 %	1,188,705
7	6	4.51	75 %	1,045,537
8	5	4.26	85 %	1,397,512
9	5	4.34	87 %	1,822,131
10	4	3.70	93 %	1,855,194
11	3	2.80	93 %	1,987,695
12	2	1.89	94 %	2,069,239
13	1	1.01	101 %	1,862,584

Tabla 5: Años que deberíamos tener de cada generación si no hubieran faltas ni deserciones y años promedio que efectivamente encontramos con panel fuzzy

Generación	Años observados	Media de años observados	% Observados	Número de niños
1	1	1.41	141 %	1,537,620
3	1	1.45	145 %	1,610,472
4	2	1.99	99 %	1,569,010
5	4	3.34	84 %	1,584,600
6	5	4.12	82 %	1,574,381
7	6	6.91	115 %	1,565,006
8	5	4.26	85 %	1,961,483
9	5	4.40	88 %	1,971,532
10	4	3.71	93 %	2,057,080
11	3	2.81	94 %	2,166,091
12	2	2.03	101 %	2,219,244
13	1	1.01	101 %	1,982,939

La Tabla 4 muestra las generaciones en renglones y los números que deberíamos tener para

cada una. Sacando un promedio simple[1] obtenemos que deberíamos tener cerca de 3 años de seguimiento por alumno, y en realidad tenemos 2.8 en el panel_ fuzzy, es decir tenemos un 93 % de los años. Esto podría explicarse con una tasa de inasistencia de 15 %. Los porcentajes más bajos corresponden a la generación 5, 6 y 7 que cursaban secundaria en el 2010.

Enlace es una prueba estandarizada que se hizo desde el 2006 al 2012 para primarias y secundarias. En primarias cubre los años de 3ro a 6to. En secundaria cubre, en algunos años, de 1ro a 3ro, en otros solo cubre 3ro, es por estos huecos que para este reporte sólo se utilizaron los datos de las pruebas Enlace de 2006 a 2012 para **primarias**.

[1]Para mayor exactitud se podría también hacer un promedio ponderado por el número de alumnos en cada generación

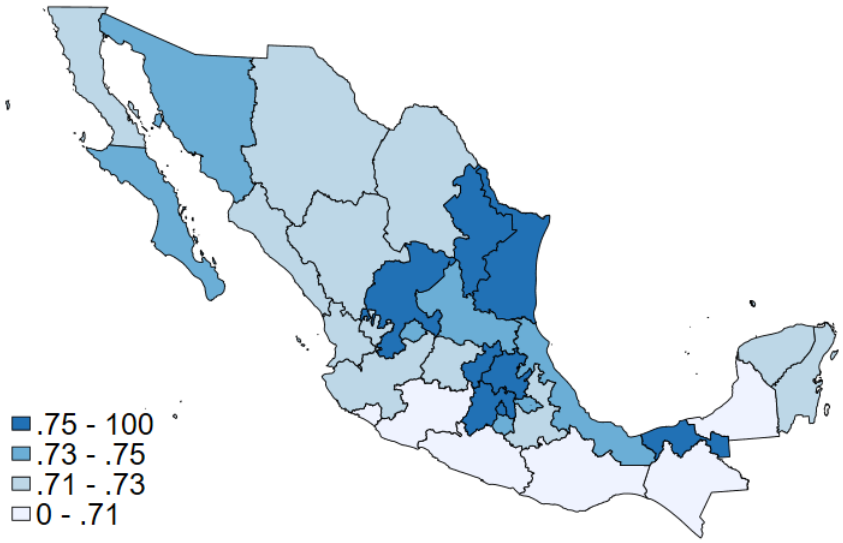
Objetivo 2: Calidad del panel de datos

En esta parte del reporte mostramos algunos estadísticos descriptivos y algunas medidas de calidad de la base de datos.

A) Calidad por estado.

Puede existir la preocupación de que las observaciones faltantes estén sesgadas. Los alumnos que faltan en la prueba deberían estar distribuidos uniformemente entre los estados. En la figura 1 se puede ver para 2009 que los estados donde una menor proporción de los alumnos presentan el exámen se encuentran en el sur de la república: Chiapas, Guerrero, Oaxaca entre otros. Esto se debe tomar en cuenta al momento de usar el panel puesto que las observaciones faltantes se concentran en ciertos estados.

Figura 1: Mapa con el ratio de Enlace entre niños inscritos en el estado



Ratio Enlace: Tomamos el número de alumnos por entidad que presentaron la prueba ENLACE en 2009 y el número de alumnos que fueron reportados como inscritos en 2009 en el formato 911 y calculamos la razón. Los colores están divididos en cuartiles y los porcentajes que representa cada cuartil es mayor cuanto más oscuro es el color.

Las tablas 6 y 7 hacen un ejercicio parecido por estado pero para todos los años. Se cuenta el número de alumnos, en la tabla 6, y el número de escuelas, en la tabla 7, que presentaron la prueba. Se puede observar que para los años 2006, 2010, 2011 y 2012 no se cuenta con observaciones de Oaxaca y para 2008 no se cuenta con observaciones de Michoacán. Esto se debe tomar en cuenta si se usan las bases para realizar un análisis podría ser buena idea excluir ambos estados.

Tabla 6: Número de alumnos que participaron en 2006-2012 por Estado

	Año						
Estado	2006	2007	2008	2009	2010	2011	2012
Aguascalientes	97,235	97,426	96,028	95,814	99,768	101,470	102,646
Baja California	48,862	225,713	232,245	229,769	243,275	249,549	249,791
Baja California Sur	15,110	41,499	38,061	44,305	46,939	48,508	49,365
Campeche	51,020	59,657	54,997	61,113	62,015	64,294	64,952
Coahuila	203,227	205,627	205,138	200,455	210,140	226,392	228,112
Colima	43,420	45,480	42,981	42,464	43,655	45,131	49,960
Chiapas	2,104	132,714	393,667	427,560	417,771	454,240	298,334
Chihuahua	222,536	248,525	254,153	260,445	263,937	260,873	259,773
Distrito Federal	579,145	599,084	584,470	595,439	602,636	605,259	599,099
Durango	118,803	117,034	99,170	124,472	125,327	130,476	133,465
Guanajuato	444,580	466,802	466,579	469,592	478,563	502,419	505,699
Guerrero	257,346	284,923	304,674	182,838	258,666	284,403	37,306
Hidalgo	190,749	203,760	209,727	202,858	216,467	215,976	221,440
Jalisco	551,085	539,870	537,296	559,672	546,973	596,841	606,678
México	402,364	1,167,650	1,166,860	1,131,111	1,206,680	1,217,535	1,229,048
Michoacán	76,683	189,079	-	123,389	140,733	165,780	73,008
Morelos	112,812	118,319	112,951	108,552	135,160	137,151	133,944
Nayarit	79,305	78,738	77,614	78,482	84,025	84,382	84,907
Nuevo León	292,875	310,380	322,168	332,165	353,411	370,452	378,857
Oaxaca	-	144,172	282,963	32,401	-	-	-
Puebla	327,083	379,480	394,825	401,633	488,263	518,159	532,271
Querétaro	147,700	147,203	144,934	147,224	157,094	159,896	159,618
Quintana Roo	93,864	92,620	94,351	92,888	95,473	101,429	104,554
San Luis Potosí	224,085	219,322	201,570	211,626	215,435	232,519	233,614
Sinaloa	80,626	220,169	211,269	204,819	203,533	215,222	216,107
Sonora	181,090	191,332	194,191	187,315	209,680	203,421	217,196
Tabasco	164,215	176,627	170,838	176,580	181,463	182,656	182,508
Tamaulipas	230,860	235,714	227,527	226,041	245,524	253,943	245,631
Tlaxcala	73,990	40,845	71,925	93,649	88,164	87,075	92,844
Veracruz	523,008	601,354	582,626	605,026	619,826	621,492	634,458
Yucatán	131,432	121,897	140,392	139,643	148,806	156,780	159,780
Zacatecas	70,538	111,810	115,677	116,914	126,477	129,305	121,686

Tabla 7: Número de escuelas que participaron en 2006-2012 por Estado

	año						
Estado	2006	2007	2008	2009	2010	2011	2012
Aguascalientes	671	713	683	690	685	693	685
Baja California	714	1,481	1,530	1,563	1,602	1,653	1,652
Baja California Sur	215	372	335	347	360	364	376
Campeche	662	675	670	674	677	681	681
Coahuila	1,696	1,717	1,717	1,744	1,758	1,779	1,783
Colima	416	440	426	427	434	438	439
Chiapas	469	3,460	5,890	6,217	5,770	6,253	3,999
Chihuahua	2,225	2,508	2,417	2,469	2,472	2,472	2,461
Distrito Federal	3,218	3,344	3,337	3,314	3,310	3,266	3,242
Durango	2,010	1,984	2,035	2,069	2,067	2,064	2,078
Guanajuato	4,585	4,649	4,623	4,631	4,657	4,689	4,313
Guerrero	4,087	3,874	3,881	2,436	3,219	3,484	606
Hidalgo	3,073	3,118	2,706	2,723	2,722	2,710	2,716
Jalisco	5,364	5,704	5,714	5,782	5,801	5,817	5,447
México	3,337	7,248	7,585	7,603	7,690	7,706	7,452
Michoacán	1,528	2,566	-	2,037	2,127	2,331	977
Morelos	949	976	977	933	1,017	1,027	1,017
Nayarit	982	1,130	1,139	1,144	1,166	1,177	983
Nuevo León	2,405	2,481	2,505	2,547	2,602	2,655	2,689
Oaxaca	-	3,557	4,175	646	-	-	-
Puebla	3,577	4,390	3,983	3,477	4,114	4,197	4,190
Querétaro	1,188	1,319	1,197	1,222	1,236	1,241	1,242
Quintana Roo	710	718	734	745	759	778	779
San Luis Potosí	3,159	2,930	2,758	2,747	2,747	2,756	2,772
Sinaloa	1,477	2,407	2,303	2,321	2,336	2,372	2,331
Sonora	1,667	1,699	1,731	1,730	1,747	1,617	1,761
Tabasco	1,868	2,120	1,923	1,927	1,924	1,922	1,920
Tamaulipas	2,250	2,335	2,246	2,241	2,271	2,319	2,318
Tlaxcala	643	385	622	672	760	718	691
Veracruz	8,594	9,460	8,593	8,627	8,630	8,565	8,670
Yucatán	1,324	1,221	1,217	1,228	1,254	1,270	1,280
Zacatecas	1,551	1,787	1,791	1,992	1,778	1,774	1,705

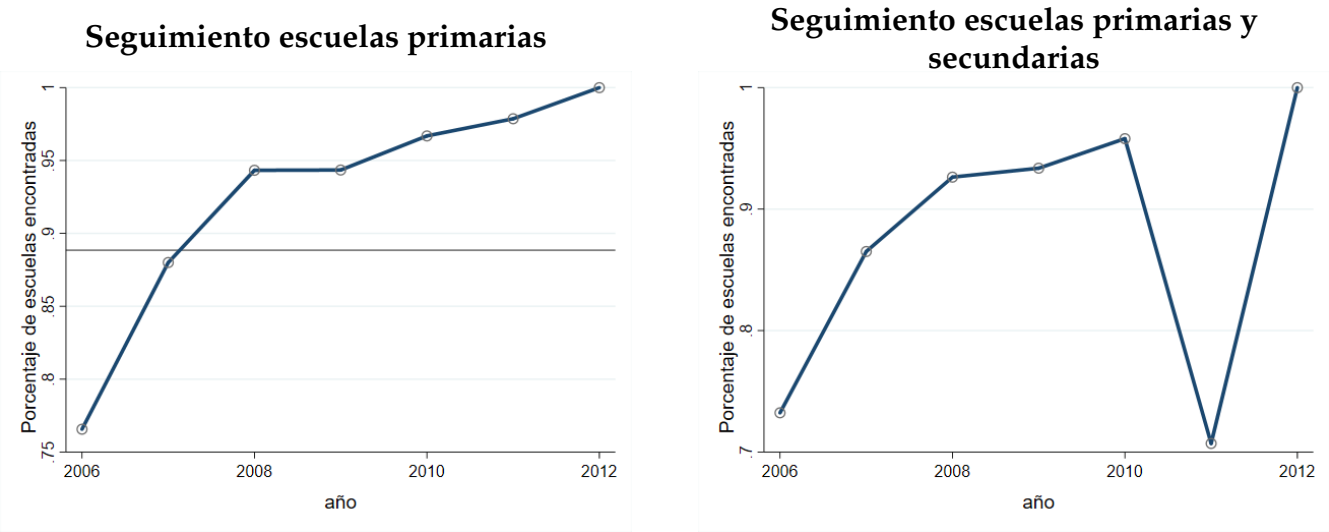
Como se puede observar en las páginas anteriores, no contamos datos para todos los años para los estados de Oaxaca y Michoacán es por eso que el resto del análisis se hace omitiendo ambos estados.

B) Seguimiento.

Para esta sección se hace un seguimiento en retrospectiva de las escuelas y de los alumnos en sus respectivas generaciones.

En la Figura 2 se toman todas las escuelas que se observan en 2012 y se buscan en años anteriores. En 2006 se encuentran

Figura 2: Porcentaje de escuelas encontradas de forma retrospectiva.



Nota: Tomamos a las escuelas que aparecen en 2012 e hicimos un seguimiento de ellas en el pasado: hasta 2006 encontramos alrededor del 77 %. La caída en 2011 se debe a la falta de información sobre escuelas secundarias en 2011.

La Tabla 2.1 muestra en los renglones los años escolares desde tercero de primaria hasta tercero de secundaria, y en las columnas los años calendario para los que contamos con Enlace: 2006 a 2012. Los colores representan la misma generación.

Lo que se observa aquí es que para algunas generaciones podemos seguir a los alumnos por varios años. Por ejemplo, la generación 9 corre desde 3ro de primaria en 2008 hasta 1ro de secundaria en 2012. Es decir, con el panel de datos que formamos pueden seguirse a 6 generaciones a través del tiempo por al menos 3 años [2]. Cabe notar que, en 2006, 2007 y 2008, segundo

[2]Como no usamos las pruebas PLANEA son solo 6 las generaciones que podemos seguir en el tiempo por al menos tres años.

Tabla 2.1: Versión recortada de la tabla 2 con los años de interés.

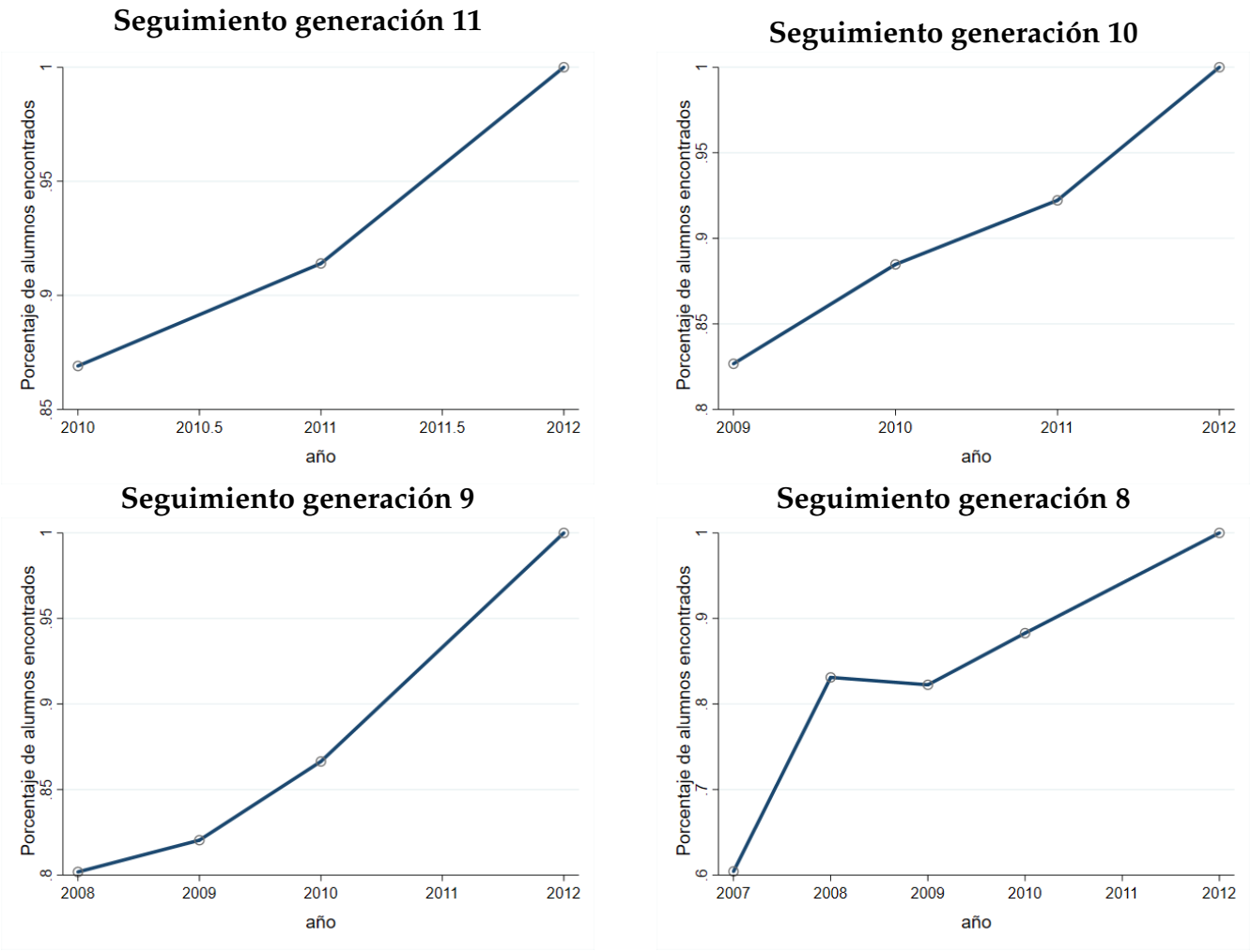
grado	2012	2011	2010	2009	2008	2007	2006
3ro prim	13	12	11	10	9	8	7
4to prim	12	11	10	9	8	7	6
5to prim	11	10	9	8	7	6	5
6to prim	10	9	8	7	6	5	4
1ro secu	9		7	6			
2do secu	8		6	5			
3ro secu	7		5	4	3	2	1

y primero de secundaria no presentaron la prueba Enlace y que la base que nos proporcionaron de 2011 solo tiene primaria.

A diferencia de la Figura 2 que busca a las escuelas en años pasados, la Figura 3 hace un seguimiento retrospectivo de una generación en específico a nivel alumno. Esto añade calidad al análisis porque permite identificar si los datos faltantes se dan en una generación en especial. Además, hay generaciones que en ciertos años no presentaron examen, como primero de secundaria en 2007: el seguimiento por generación toma estos casos en cuenta y los omite.

La Figura 3 se lee de la siguiente manera. Si el título de la gráfica es "seguimiento generación 11" tomamos a los niños de la generación 11 que iban en quinto de primaria en 2012 y los buscamos en años anteriores. Por ejemplo, en 2011 encontramos únicamente al 92 % de los niños que iban en quinto de primaria en 2012.

Figura 3: A cuántos alumnos encontramos retrospectivamente a partir del último año observado.



La Figura 3 muestra que es posible encontrar a una proporción importante de los niños de secundaria en los años anteriores: cerca de 80 % cinco años antes, y cerca de 87 % tres años antes [3].

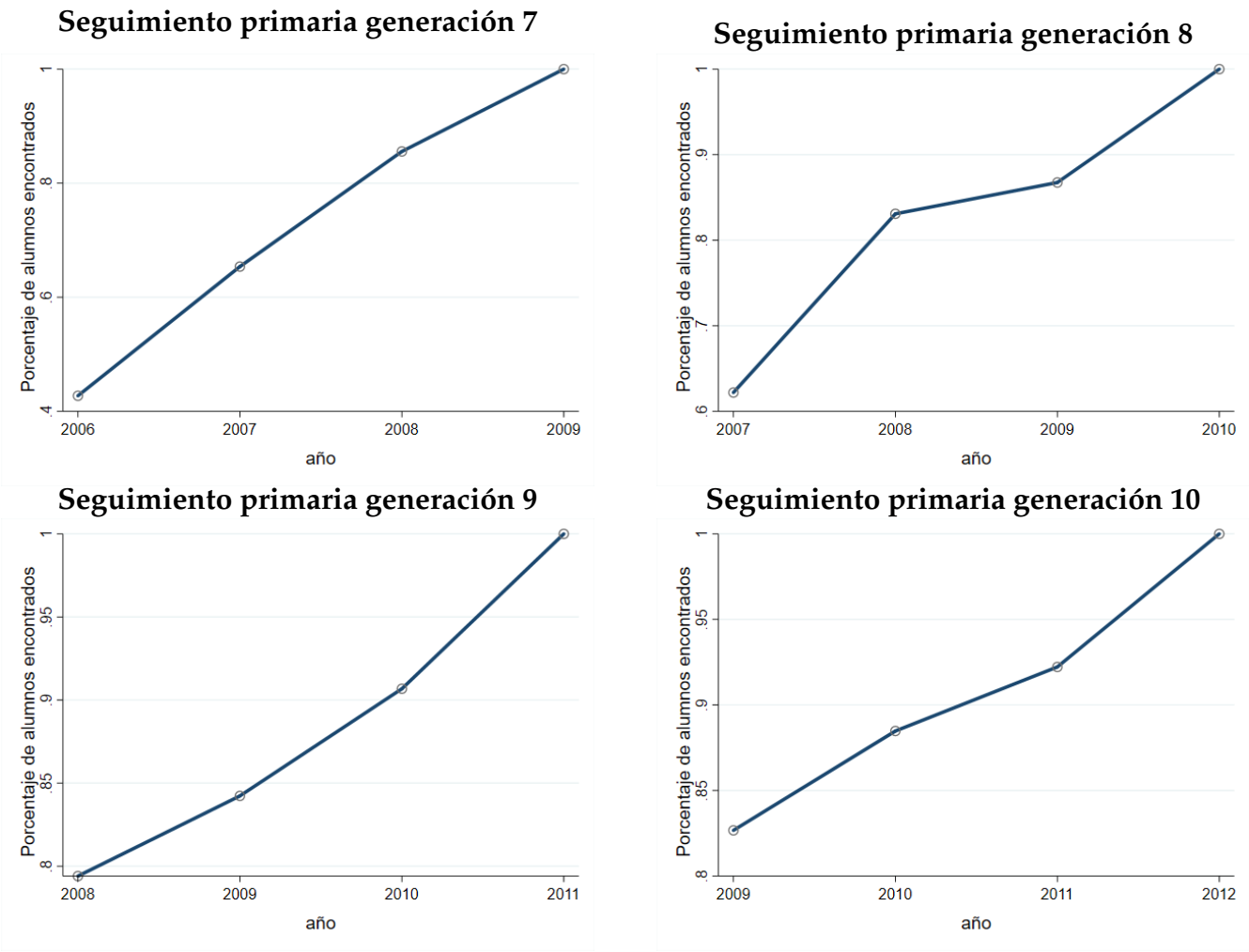
La Figura 4a toma las cuatro generaciones para las que observamos toda la primaria y, empezando de 6to de primaria, se pregunta qué fracción de personas se encuentran en años escolares anteriores.

Las gráficas de la figura 4a muestran una pendiente positiva esto indica que de los alumnos que iban en sexto, en un año dado, se encuentra un porcentaje menor el año anterior. La progresión monótona no decreciente no es explicada por deserción dado que estamos condicionando en

[3] En el panel izquierdo superior, solo observamos quinto cuarto y tercero de primaria, por eso hay menos puntos

que tenemos información de la escuela a la que asistieron en los últimos años. Una posibilidad es que se encuentren menos alumnos en los años anteriores porque con el tiempo se incorporaron más escuelas a la prueba Enlace y se obtuvieron datos de más alumnos por escuela. Es decir, los estudiantes que estaban en sexto de primaria y no se encontraron en los años anteriores pudieron haber faltado; su escuela pudo no participar en la prueba antes de ese año; los datos de su escuela en los años anteriores estaban incompletos o no hubo forma de seguir al estudiante a través de los años por errores al registrar su nombre o CURP.

Figura 4a: A cuántos alumnos encontramos retrospectivamente desde sexto de primaria empezando en distintos años



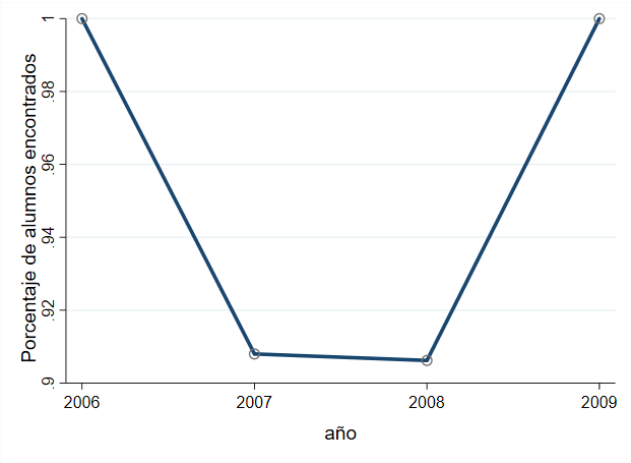
Nota: La Figura 4 se concentra en alumnos de primaria y muestra que dependiendo de la generación, partiendo de sexto de primaria, podemos encontrar entre 70 % y 83 % de los niños en 3ro de primaria.

La Figura 4b toma a los alumnos de las cuatro generaciones para las que observamos toda la primaria y que observamos tanto en tercero como en sexto de primaria y los busca en los años

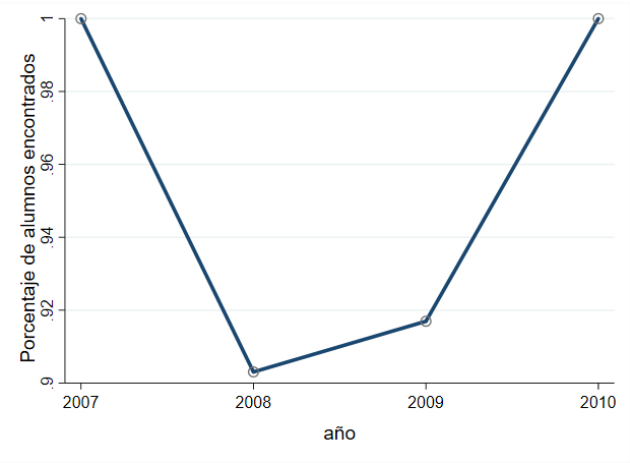
intermedios. Se escuentra alrededor del 90 % de los alumnos.

Figura 4b: A cuántos alumnos encontramos retrospectivamente a partir del último año observado condicionando en que tenemos observaciones de los alumnos en tercero y sexto de primaria.

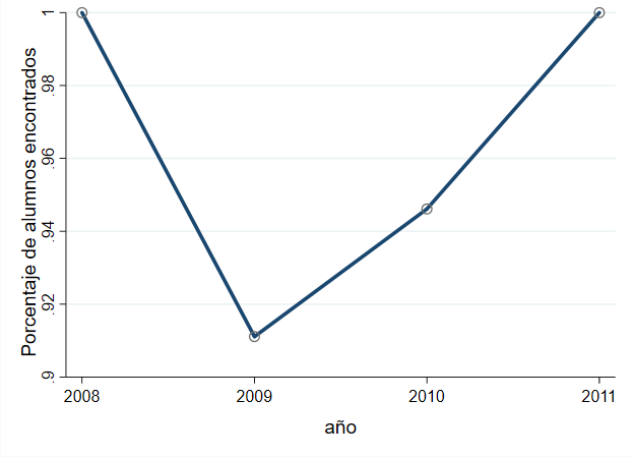
Seguimiento primaria condicionada generación 7



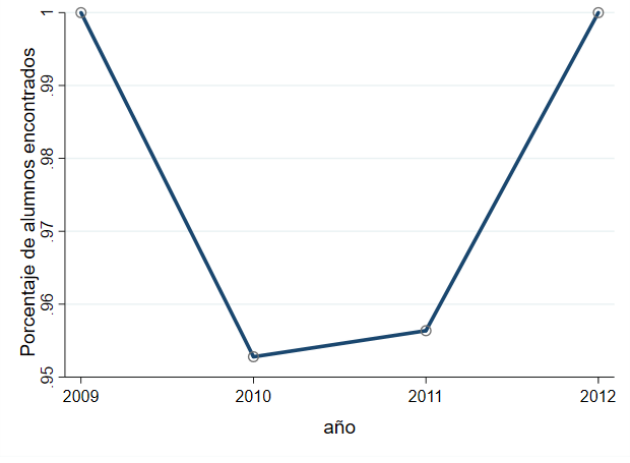
Seguimiento primaria condicionada generación 8



Seguimiento primaria condicionada generación 9



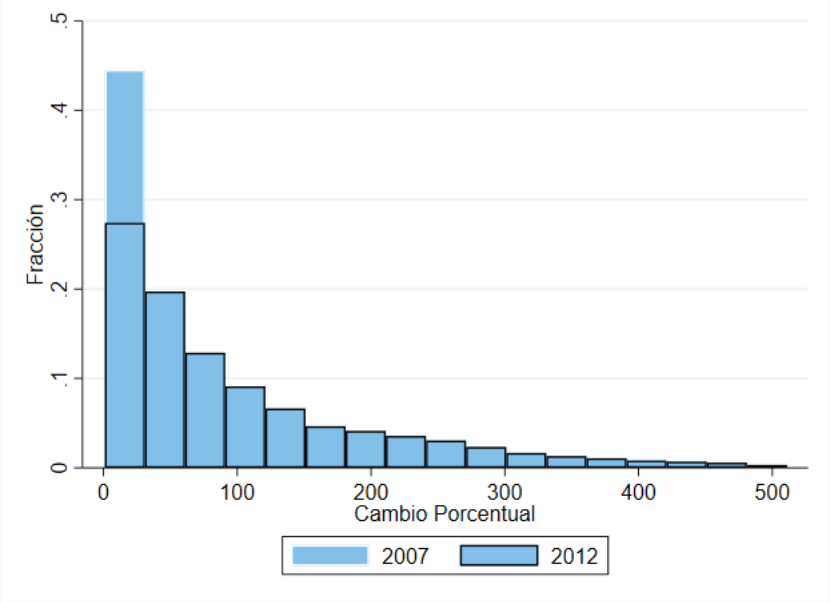
Seguimiento primaria condicionada generación 10



C) Tamaño de matricula

Otra forma de explorar la calidad de la base de datos es por medio de la exploración del tamaño de la escuela medido por el número de personas que hacen la prueba Enlace. La figura 4 muestra un histograma del número de estudiantes por escuela que hizo la prueba Enlace en el 2007 y en el 2012. Como puede apreciarse, existe un mayor número de escuelas con menos de 50 estudiantes en el 2007 comparado con el 2012. En otras palabras, en el 2012 es mayor el número de alumnos por escuela en comparación con el 2007. El número de alumnos promedio por escuela es 84 en el 2007 y 118 en el 2012.

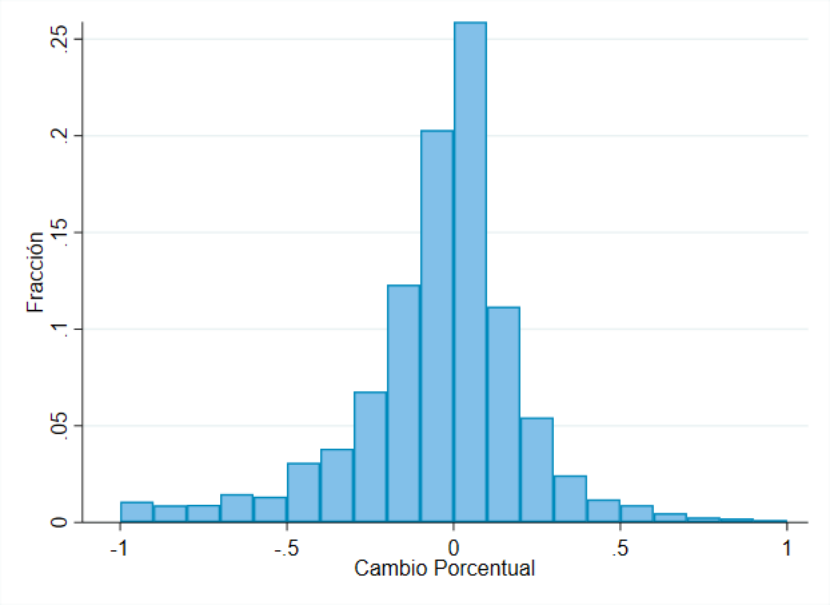
Figura 5: Histograma del tamaño de las escuelas en promedio en 2007 y 2012.



Nota: El tamaño de la escuela se encontró agrupando por año y por CCT. En el 2012 participaron secundarias y primarias mientras que en el 2007 solo primarias, sin embargo, se contaron como escuelas separadas ya que tienen un CCT diferente

La Figura 6 compara el tamaño de las escuelas entre años subsecuentes. En particular, para cada escuela calcula el cambio porcentual entre un año y otro. Esperaríamos que la gran mayoría de las escuelas mantuviera, aproximadamente, su mismo tamaño a lo largo de los años. Podemos apreciar que cerca del 30 % de las escuelas reduce su asistencia en 1 %, mientras que el 20 % de las escuelas lo aumenta en 1 %. Muy pocas escuelas cambian el tamaño de su matrícula en más de 5 %, lo cual es un indicativo de la calidad de los datos.

Figura 6: Variación en asistencias: la misma escuela a través del tiempo



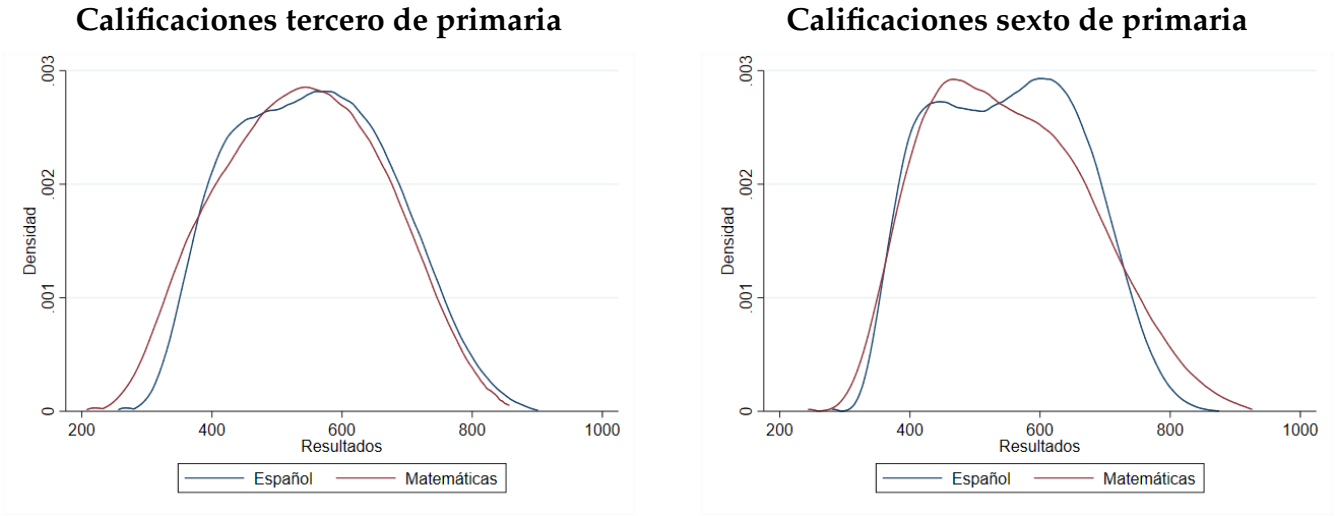
Nota: La variación se calculó como el cambio porcentual del tamaño de la escuela entre años consecutivos. Por cada año, para cada escuela se calculó el número de alumnos que participaron en la prueba y se comparó con el año siguiente para encontrar el cambio porcentual

D) Distribución de calificaciones y estabilidad en el tiempo

Distribución

Otra forma de medir la calidad de datos es con el contenido de las calificaciones. La calidad de la educación es difícil de cambiar y por tanto debe exhibir inercia tanto a nivel escuela, como a nivel alumno. La Figura 6 muestra la distribución de los resultados de la prueba en tercero de primaria y en sexto de primaria en el 2011. Los resultados están separados por asignatura: la línea roja muestra los resultados de Matemáticas y la azul los de Español.

Figura 7: Distribución de calificaciones en Matemáticas y Español para 3ro de primaria y 6to de primaria en el año 2011.



Nota: Se muestra a la distribución del 2011 como caracterización de los otros años. Es decir, los resultados se comportan de forma muy parecida en los otros años y los otros grados

La Figura 7 muestra el cambio porcentual en la calificación promedio de los alumnos de una escuela. Es decir, primero se obtiene el promedio dentro de una escuela en cada año y después se calcula cuánto cambia este promedio en años consecutivos. Se usaron todos los años para los cuales existe la escuela (primarias y secundarias). El resultado muestra estabilidad: la gran mayoría de las escuelas cambian menos del 1 % la calificación.

Estabilidad en el tiempo a nivel alumno

Es posible hacer la estimación de estabilidad a nivel estudiante. En particular, estimamos un modelo AR(1) a nivel estudiante con la siguiente ecuación: $C_{it} = \alpha + \beta C_{(i,t-1)} + \epsilon_{it}$, donde C_{it} es la calificación del alumno i en el año t.

Tabla 8: Add caption

Persistencia a nivel persona		
	Matemáticas	Español
	(1)	(2)
ß (t-1)	0.523***	
	(0.000)	
ß (t-1)		0.520***
		(0.000)
Constante	0.001***	0.004***
	(0.000)	(0.000)
Observaciones	28,221,285	28,199,079
Número de CURP	14,918,877	14,901,222

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

E) Cierres y aperturas de escuelas

Esta sección cuenta, por cada año, cuántos cierres y cuántas aperturas de escuelas hubo. Usamos dos definiciones de cierre:

- **Cierres simples**: es cuando una escuela identificada por su folio CCT está en un año y no está el siguiente año ni en ningún año subsecuente. Es decir, ningún alumno presenta la prueba Enlace en ningún año posterior.
    - En esta definición, si la misma escuela (mismo edificio y mismos maestros) cambiara de número de folio CCT lo contaríamos como un cierre de escuela. La siguiente definición es más estricta.
- **Cierres con alumnos en otras escuelas**: una escuela cierra en esta definición cuando (al igual que la definición anterior) identificada por su folio CCT está en un año y no está el siguiente año ni en ningún año subsecuente, pero adicionalmente, encontramos al menos a 25 % de los alumnos del año anterior en otras escuelas. La Tabla 4 también considera otra definición en donde se encuentran 40 % de los alumnos.

Tabla 9: Escuelas cerradas por años para primarias públicas (definiciones múltiples)

Año	Total	num. Esc. Cierran	%	25 % encontrados	%	40 % encontrados	%
2006	58,834	182	0.31 %	140	0.24 %	103	0.18 %
2007	72,194	299	0.41 %	236	0.33 %	173	0.24 %
2008	74,025	742	1.00 %	350	0.47 %	257	0.35 %
2009	70,144	399	0.57 %	347	0.49 %	290	0.41 %
2010	71,225	350	0.49 %	293	0.41 %	248	0.35 %
2011	71,868	3,409	4.74 %	1,043	1.45 %	447	0.62 %

Tabla 10: Escuelas cerradas por años para primarias privadas (definiciones múltiples)

Año	Total	num. Esc. Cierran	%	25 % encontrados	%	40 % encontrados	%
2006	5,187	69	1.33 %	58	1.12 %	37	0.71 %
2007	6,671	112	1.68 %	90	1.35 %	80	1.20 %
2008	6,945	163	2.35 %	133	1.92 %	115	1.66 %
2009	7,381	137	1.86 %	125	1.69 %	113	1.53 %
2010	7,579	140	1.85 %	123	1.62 %	107	1.41 %
2011	7,774	233	3.00 %	155	1.99 %	144	1.85 %

Aperturas

De forma análoga, pueden calcularse las aperturas de escuelas. La Tabla 5 muestra las aperturas de escuelas usando la siguiente definición:

- **Escuela que abre**: decimos que una escuela abre en un año t si el CCT existe a partir de t, pero en años anteriores a t no existía.
- **Escuela que abre ajustada**: decimos que una escuela abre ajustadamente en un año t si el CCT existe a partir de t, pero en años anteriores a t no existía, y, además, podemos encontrar al 25 % (40 %) de los alumnos de la escuela que aparece por primera vez en determinado año, un año antes, en otras escuelas. Esto nos permite excluir escuelas que probablemente no abrieron y, más bien, se incorporaron posteriormente al examen.

Tabla 4:Escuelas que se incorporan a la base, primarias privadas

Año	Total	num. esc abren	%	25 % encontrados	Porcentaje	40 % encontrados	%
2006	5,187						
2007	6,671	577	8.65 %	87	1.30 %	51	0.76 %
2008	6,945	247	3.56 %	97	1.40 %	61	0.88 %
2009	7,381	233	3.16 %	125	1.69 %	81	1.10 %
2010	7,579	255	3.36 %	148	1.95 %	98	1.29 %
2011	7,774	214	2.75 %	153	1.97 %	104	1.34 %
2012	7,826	264	3.37 %	180	2.30 %	126	1.61 %

Tabla 5: Escuelas que se incorporan a la base, primarias publicas

Año	Total	num. esc abren	%	25 % encontrados	Porcentaje	40 % encontrados	%
2006	58,834						
2007	72,194	2161	2.99 %	188	0.26 %	125	0.17 %
2008	74,025	560	0.76 %	225	0.30 %	127	0.17 %
2009	70,144	413	0.59 %	306	0.44 %	257	0.37 %
2010	71,225	352	0.49 %	242	0.34 %	185	0.26 %
2011	71,868	468	0.65 %	336	0.47 %	265	0.37 %
2012	64,910	329	0.51 %	248	0.38 %	211	0.33 %

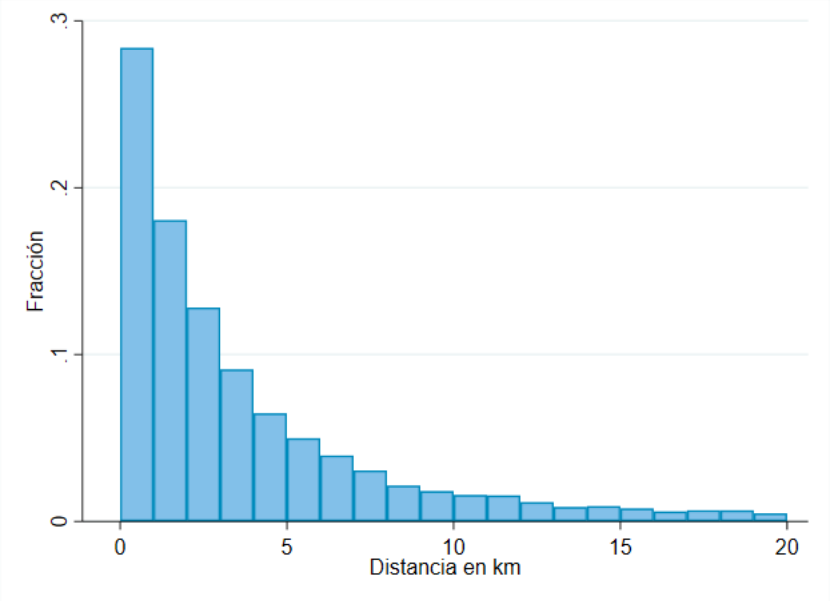
F) Distancias

Aunque no es parte de la información proporcionada por el Banco Mundial, hicimos un breve análisis usando distancia entre escuelas. Para calcular estas distancias primero conseguimos información adicional de las coordenadas GPS de cada una de las escuelas. Nos concentramos en distancias por calle (i.e. no lineales entre escuelas).

Otro chequeo de información que hicimos tiene que ver con las distancias entre una escuela que cierra y la escuela que recibe a los alumnos se cambian. Después, habiendo identificado las escuelas que cerraron, seguimos a los niños que estaban en estas escuelas a las nuevas escuelas. Podemos encontrar al 49 % de los alumnos de primaria en escuelas que cierran en una escuela diferente el siguiente año. Paso seguido, calculamos las distancias por caminos (*road distance*) entre esas escuelas en una base de datos donde una observación es un niño que se cambia de escuela.

La Figura 9 muestra las distancias entre la escuela que cerró en el año t y la escuela en la que encontramos a los alumnos en el año $t+1$. Que las distancias sean razonables, con un promedio menor a 2km, sugiere que la información de cierre de las escuelas y donde los encontramos después (y la localización GPS) es de buena calidad.

Figura 9: Distancias entre escuelas cerradas y las escuelas reemplazo



Nota: No consideramos los casos en los cuáles la escuela cambió de CCT y la distancia entre las escuelas era cero.

Objetivo 3: Análisis comparativo usando panel exacto Enlace y panel 911

El panel de datos 911 contiene las variables más importantes que se capturan en las encuestas que las escuelas de México hacen tanto al inicio del ciclo escolar como al termino de este. En este panel en particular, solo se cuenta con las variables y observaciones que corresponden al inicio del ciclo escolar.

En las tablas 13, 14 y 15 se muestran los datos descriptivos del panel 911.

Tabla 13: Panel de datos armado por nosotros

Nombre	Número de CCT	Número de observaciones	Variables
panel_911	84,658	859,808	184

Tabla 14: Número de alumnos inscritos, registrados en el 911 por estado y año.

	Año						
Estados	2006	2007	2008	2009	2010	2011	2012
Aguascalientes	101,613	100,772	101,059	102,908	105,878	106,631	106,334
Baja California	239,574	244,708	247,704	250,885	267,194	266,944	265,110
Baja California Sur	43,861	45,033	46,638	48,734	52,913	54,011	54,569
Campeche	65,981	65,563	64,765	64,572	67,710	68,523	68,185
Chiapas	220,551	220,038	219,469	222,412	245,087	248,429	247,597
Chihuahua	49,968	47,679	45,422	45,301	48,179	52,630	53,673
Coahuila	336,738	335,326	332,422	323,520	328,450	331,528	332,532
Colima	266,135	268,976	274,806	279,090	279,012	277,356	275,271
Distrito Federal	649,853	627,995	607,622	616,730	633,043	638,324	625,577
Durango	138,636	137,739	137,373	135,872	142,722	143,938	144,038
Guanajuato	480,331	484,094	485,283	488,909	523,010	532,454	527,752
Guerrero	293,258	286,457	277,905	271,355	273,370	272,325	268,807
Hidalgo	201,449	197,233	193,281	194,627	203,215	207,342	209,523
Jalisco	585,603	577,697	573,870	594,528	643,619	656,179	643,980
Michoacán	1,227,642	1,217,839	1,211,787	1,229,181	1,270,837	1,283,076	1,286,140
Morelos	377,192	373,784	362,562	362,562	348,791	345,702	346,471
México	145,199	142,321	139,574	137,596	149,652	150,791	150,660
Nayarit	80,212	79,191	77,769	78,515	85,028	85,823	86,285
Nuevo León	327,296	338,362	343,180	354,585	391,174	395,812	392,916
Oaxaca	290,427	290,427	267,250	265,570	266,659	266,604	264,194
Puebla	477,322	472,885	468,498	470,938	493,538	513,053	517,182
Querétaro	148,692	147,872	147,305	149,342	162,216	165,473	165,080
Quintana Roo	100,197	101,596	100,363	100,588	105,482	110,280	112,925
San Luis Potosí	220,106	215,210	211,085	211,106	226,052	228,635	226,680
Sinaloa	236,266	237,645	226,606	224,555	227,593	229,246	225,938
Sonora	202,760	202,335	203,124	207,070	229,863	230,296	228,790
Tabasco	184,904	181,664	178,344	182,122	190,202	190,870	187,604
Tamaulipas	255,114	252,686	250,567	258,369	270,447	270,531	265,056
Tlaxcala	103,882	103,537	102,484	100,568	100,771	101,967	103,779
Veracruz	608,913	601,926	590,226	587,652	608,299	605,541	595,255
Yucatán	145,386	140,915	138,858	140,409	154,569	156,986	6,643
Zacatecas	130,478	129,590	125,580	126,817	137,564	137,616	135,948

Tabla 15: Número de escuelas registradas en el 911 en 2006-2012 por Estado

	Año						
Estados	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Aguascalientes	675	682	688	700	698	686	680
Baja California	1,464	1,496	1,522	1,570	1,613	1,626	1,614
Baja California Sur	344	355	363	369	385	389	392
Campeche	625	626	623	629	631	635	636
Chiapas	1,719	1,722	1,745	1,779	1,801	1,800	1,789
Chihuahua	433	435	437	441	454	462	465
Coahuila	3,526	3,524	3,535	3,549	3,558	3,577	3,583
Colima	2,142	2,142	2,159	2,184	2,171	2,180	2,170
Distrito Federal	3,441	3,392	3,424	3,419	3,311	3,333	3,296
Durango	1,866	1,883	1,883	1,879	1,874	1,880	1,885
Guanajuato	4,344	4,373	4,352	4,387	4,373	4,380	4,376
Guerrero	3,127	3,133	3,140	3,106	3,107	3,110	3,113
Hidalgo	2,123	2,127	2,134	2,138	2,150	2,127	2,150
Jalisco	5,304	5,355	5,368	5,392	5,435	5,450	5,477
Michoacán	7,086	7,182	7,208	7,238	7,304	7,374	7,433
Morelos	4,395	4,417	4,405	4,405	4,458	4,488	4,503
México	972	993	1,011	1,015	1,054	1,076	1,097
Nayarit	834	844	884	889	907	885	885
Nuevo León	2,477	2,508	2,533	2,600	2,691	2,720	2,754
Oaxaca	3,078	3,078	3,102	3,109	3,141	3,153	3,148
Puebla	3,460	3,473	3,477	3,485	3,501	3,512	3,507
Querétaro	1,125	1,133	1,145	1,159	1,172	1,185	1,189
Quintana Roo	629	636	659	683	707	712	717
San Luis Potosí	2,458	2,435	2,431	2,436	2,434	2,442	2,351
Sinaloa	2,328	2,319	2,326	2,330	2,374	2,351	2,290
Sonora	1,663	1,680	1,694	1,679	1,701	1,707	1,702
Tabasco	1,836	1,835	1,836	1,836	1,837	1,836	1,831
Tamaulipas	2,211	2,266	2,294	2,338	2,375	2,397	2,386
Tlaxcala	663	670	672	686	700	705	708
Veracruz	7,751	7,767	7,770	7,761	7,719	7,739	7,771
Yucatán	1,065	1,068	1,066	1,089	1,101	1,114	1,125
Zacatecas	1,825	1,813	1,812	1,803	1,790	1,782	1,760

Para contar con los datos necesarios para el análisis que realizamos entre en el panel de datos exacto de Enlace y el panel 911 hicimos lo siguiente:

Primero, contamos el total de alumnos en ambas bases de datos (Se debe prestar atención al hecho de que la base panel está conformada a nivel escuela (CCT), mientras que la base Enlace está a nivel individual (CURP)).

Después, contabilizamos el número de estudiantes atrasados en el panel de Enlace usando el criterio de la SEP para estudiantes atrasados. Según la SEP, típicamente un estudiante empieza la primaria entre los 6 y 7 años y la termina entre los 11 y 12 años. Utilizando la información del CURP de los estudiantes y su grado escolar en un determinado año pudimos determinar si un alumno estaba atrasado; por ejemplo, si en particular un alumno de 10 años se encontraba en 2009 cursando 3ero de primaria, entonces lo clasificábamos como un estudiante atrasado pues a esa edad debería estar estudiando 4to o 5to de primaria.

Una vez contados los alumnos atrasados en el panel Enlace, contamos los alumnos repetidores. El criterio para clasificar a un repetidor es sencillo; si en la base de datos un alumno ha cursado el mismo grado, pero en distintos años, entonces se le considera como un repetidor o un alumno reprobado.

Finalmente, contabilizamos este tipo de observaciones, pero en el panel 911. La diferencia con respecto al panel Enlace es que el panel 911 contiene explícitamente la información de alumnos atrasados y alumnos repetidores.

Para saber si la base de datos Enlace estaba sesgada se hizo, con los tres tipos de observaciones que mencionamos arriba, una razón entre el número de observaciones contabilizadas en la base Enlace y la base 911. Idealmente se esperaría que la razón entre alumnos contados en la base Enlace y alumnos contados en la base 911 fuera igual a uno, no obstante, esto no ocurrió así y los siguientes histogramas lo muestran.

Figura 10: Razón entre el número de alumnos que presentaron la prueba ENLACE y el número de alumnos registrados en el formato 911, por escuela y año.

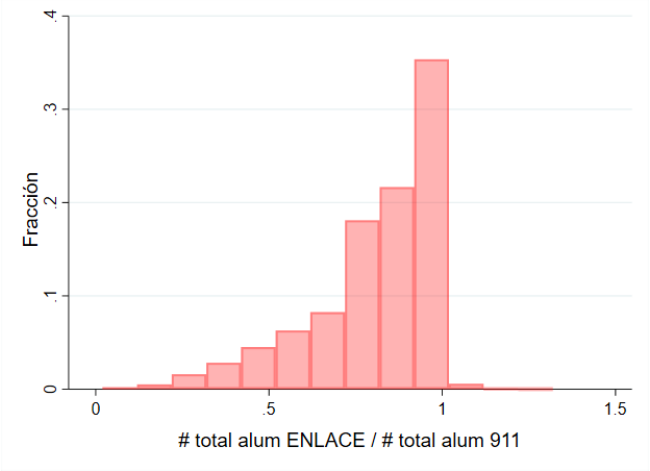


Figura 11: Razón entre el número de alumnos reprobados que presentaron la prueba ENLACE y el número de alumnos reprobados registrados en el formato 911, por escuela y año.

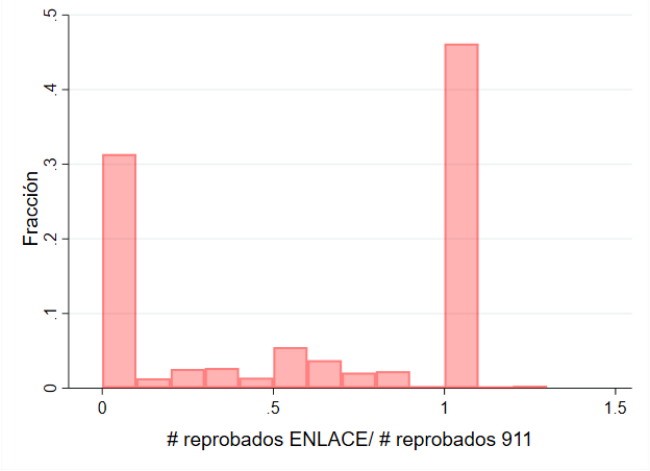
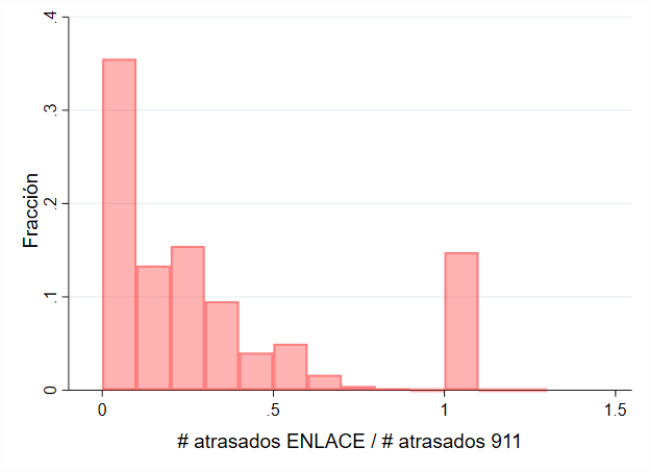


Figura 12: Razón entre el número de alumnos atrasados que presentaron la prueba ENLACE y el número de alumnos atrasados registrados en el formato 911, por escuela y año.



Nota: El número de reprobados en la prueba ENLACE se calculó observando qué alumnos presentaban la prueba en el mismo grado en dos años consecutivos. El número de alumnos atrasados se calculó utilizando la fecha de nacimiento que se obtuvo del CURP de los alumnos. Un niño está atrasado si tiene 10 o más años en tercero de primaria y así para los demás grados.

En el histograma de la figura 10, podemos observar que la forma corresponde con el porcentaje de inasistencia en la prueba Enlace (al rededor de 15 %). Sin embargo, las formas de los histogramas de las figuras 11 y 12 son menos intuitivas.

En la figura 11 podemos observar que las razones que están cercanas a cero y a uno tienen una gran acumulación de observaciones. Esto levanta la sospecha de que podría haber un gran número de escuelas, cuya razón es cercana al cero, que intentan disuadir a los alumnos reprobados a presentar la prueba Enlace o que simplemente dichos alumnos tienen mayor apatía relativa por presentar la pruebe Enlace; mientras que también podría existir un gran número de escuelas, cuya razón es cercana a uno, que no intervienen en la asistencia a la prueba de sus alumnos reprobados.

Asimismo, observamos una tendencia similar, pero no tan marcada, para el histograma de la figura 12 que corresponde a los alumnos atrasados.

A continuación, se presenta en la tabla 16 algunos estadísticos descriptivos de los tres tipos de ratios que obtuvimos.

Tabla 16: Ratio entre pruebas ENLACE y el formato 911

Variables	Ratio reprobados	Ratio atrasados	Total
Num. Observaciones	269,607	269,607	269,607
Media	0.581	0.298	0.807
Desviación e.	0.446	0.337	0.185
p1	0	0	0.251
p5	0	0	0.415
p10	0	0	0.525
p25	0	0	0.725
p50	0.750	0.200	0.857
p75	1	0.400	0.950
p90	1	1	0.986
p95	1	1	1
p99	1	1	1