

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

REPORTE PARA EL BANCO MUNDIAL

Bernardo García Bulle

Noviembre 2018

Objetivo

Este reporte tiene los siguientes objetivos:

1. Formar un panel de datos 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 matches*): 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.

Entregables

Objetivo 1: Bases de datos

1. Base de datos panel Enlace con match de CURP *exactos*
2. Base de datos panel Enlace con match de CURP *exactos + fuzzy*
3. Do file de Stata que crea ambas bases de datos

Objetivo 2: Calidad del panel y los datos

1. Este reporte
2. Do file que hace el análisis y graficas de este reporte.

Es decir: se entrega este reporte, dos bases de datos y dos do-files.

Objetivo 1: Generación de las bases de datos panel

Partimos de 18 bases de datos entregadas por el Banco Mundial con los siguientes nombres:

Tabla 1: Bases de datos entregadas por el Banco Mundial

Nombre	Número de Folios	Número de CURP de 18 dígitos	Número de CURP de 16 o más dígitos	Número de CCT	Número de variables	Tamaño en memoria (KB)
ENLACE2006 (1)	9,529,490	-	-	111,316	15	969,118
enl06nal_nombres	9,218,490	7,460,501	8,377,058	-	6	999,271
enl07_A	3,966,280	-	-	45,876	20	547,981
enl07_B	6,182,386	-	-	74,020	20	857,540
enl07nal_nombres	10,148,666	9,507,138	10,860,888	-	6	1,139,745
RESULT_ALUMNOS_08_A	4,306,540	-	-	51,539	21	407,949
enl08_B	5,646,800	-	-	68,433	23	842,895
enl08nal_nombres	9,910,885	9,719,469	10,130,905	-	7	909,789
RESULT_ALUMNOS_09_A	8,029,920	-	-	88,285	30	846,912
RESULT_ALUMNOS_09_B	5,157,768	-	-	29496	32	946,774
enl09nal_nombres	13,187,682	12,735,721	13,315,650	-	8	1,455,283
RESULT_ALUMNOS_10_A	6,054,266	-	-	52,526	8	266,059
RESULT_ALUMNOS_10_B	6,054,266	-	-	67,379	8	278,878
RES_ENLACE_10_2.csv ¹	13,772,359	-	-	119,905	30	2,495,186
enl10nal_nombres	-	13,537,621	14,004,223	-	3	1,156,664
resul_enlace_11	8,759,180	-	-	90,538	33	1,152,000
alumnos_curp_11	8,758,989	8,638,956	8,986,664	-	3	480,000
resul_alum_eb12	13,507,167	-	-	11,4346	32	1,411,401
nombres_enlb_12_nac	13,507,167	13,307,167	13,711,727	-	8	1,925,829

Notas: 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.

Nótese que la base de Enlace del 2011 está incompleta y la de 2010 tenía solamente cerca de la mitad de las calificaciones. Conseguimos estos 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.

¹ La base "RES_ENLACE_10_2.csv" tenía errores y nosotros la completamos.

La Tabla 2 muestra los datos que tenemos y numera las “generaciones” que podemos construir en cada año.

Tabla 2: Generaciones y años calendario

	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

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

Por ejemplo, la generación 6 puede seguirse desde cuarto de primaria hasta segundo de secundaria, mientras que la generación 12 solo puede seguirse de tercero a cuarto de primaria. Para las generaciones que llamamos 1, 2, 3 y 13 solo observamos un año.

Nótese que para 2011 no contamos con información de secundarias. Nótese, además, que en 2006, 2007 y 2008 no contamos con información de secundaria.

Aunque se pueden seguir generaciones, no necesariamente pueden seguirse todos los estudiantes. Hay varias razones para ello:

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”.

El do file adjunto a este entregable llamado **“MainDoFile.do”** toma estas 18 bases y genera las siguientes dos bases de datos panel que se describen en la Tabla 3.

Tabla 3: Paneles de datos armados por nosotros

Nombre	Número de CURP únicos	Número de CCT únicos	Número de observaciones	Número de variables	Núm. personas con los 4 años de primaria completos	Núm. personas con 3 años de secundaria completos	Promedio de años de seguimiento por alumno
Panel_exacto	28,406,734	131,109	70,584,838	11	15,069,740 (29%)	11,940,963 (23%)	2.48
Panel_fuzzy	28,527,479	131,683	74,998,313	11	17,911,584 (32%)	12,209,685 (22%)	2.62

Nota: La principal diferencia entre el *panel_exacto* y el *panel_fuzzy* es que el segundo solo considera los primeros 16 caracteres del CURP para hacer el match.

Para hacer el *Panel_fuzzy* se hizo lo siguiente: convertimos todos los CURP a mayúsculas, eliminamos todos los caracteres que no fueran ni números ni letras, usando nombres y las reglas de cómo se forman los CURP, identificamos algunos que no cumplían con las reglas de nombre –menos de 1% de los CURP-- y los arreglamos usando los nombres. Finalmente, 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* CURP de manera errónea, aun así, son pocos los que se recuperan. Pasamos de 2.48 años en promedio por alumno a 2.62.

Tabla 4: Años que deberíamos tener de cada generación si no hubiera faltas ni deserción y años promedio que efectivamente encontramos

Generación	Número de años de participación por generación	Número de años promedio de los alumnos de esa generación	Porcentaje
1	1	1.00	100%
2	1	1.00	100%
3	1	1.00	100%
4	2	1.46	73%
5	4	2.67	67%
6	5	3.32	66%
7	6	3.88	65%
8	5	3.87	77%
9	5	4.10	82%
10	4	3.52	88%
11	3	2.70	90%
12	2	1.84	92%
13	1	1.00	100%

Nota: El número de años de participación por generación se obtuvo con la tabla dos, mientras que el número de años promedio de los alumnos se obtuvo etiquetando a los alumnos por generación, contando las ocurrencias de un alumno en el panel y sacando el promedio por generación

Tabla 4.1. Años que deberíamos tener de cada generación si no hubiera faltas ni deserción y años promedio que efectivamente encontramos en primaria

<i>Generación</i>	<i>Número de años de participación por generación</i>	<i>Número de años promedio de los alumnos de esa generación</i>	<i>Porcentaje</i>
4	1	1.00	100%
5	2	1.50	75%
6	3	2.12	71%
7	4	2.76	69%
8	4	3.24	81%
9	4	3.45	86%
10	4	3.53	88%
11	3	2.71	90%
12	2	1.85	92%
13	1	1.01	101%

Notas: Para identificar a los alumnos de primaria se utilizó el identificador del CCT (clave del centro de trabajo). El campo del identificador (cuarto y quinto carácter del CCT) clasifica los diferentes tipos y niveles, y corresponde como sigue: Escuela de Educación Preescolar Indígena (CC), Escuela de Educación Preescolar (JN), Escuela de Educación Primaria Indígena (PB), Escuela de Educación Primaria (PR), Escuela de Educación Secundaria General (ES), Escuela de Educación Secundaria Técnica (ST), Telesecundaria (TV).

Tabla 4.2. Años que deberíamos tener de cada generación si no hubiera faltas ni deserción y años promedio que efectivamente encontramos en secundaria

<i>Generación</i>	<i>Número de años de participación por generación</i>	<i>Número de años promedio de los alumnos de esa generación</i>	<i>Porcentaje</i>
1	1	1.00	100%
2	1	1.00	100%
3	1	1.01	101%
4	1	1.00	100%
5	2	1.84	92%
6	2	1.83	91%
7	2	1.78	89%
8	1	1.00	100%
9	1	1.00	100%

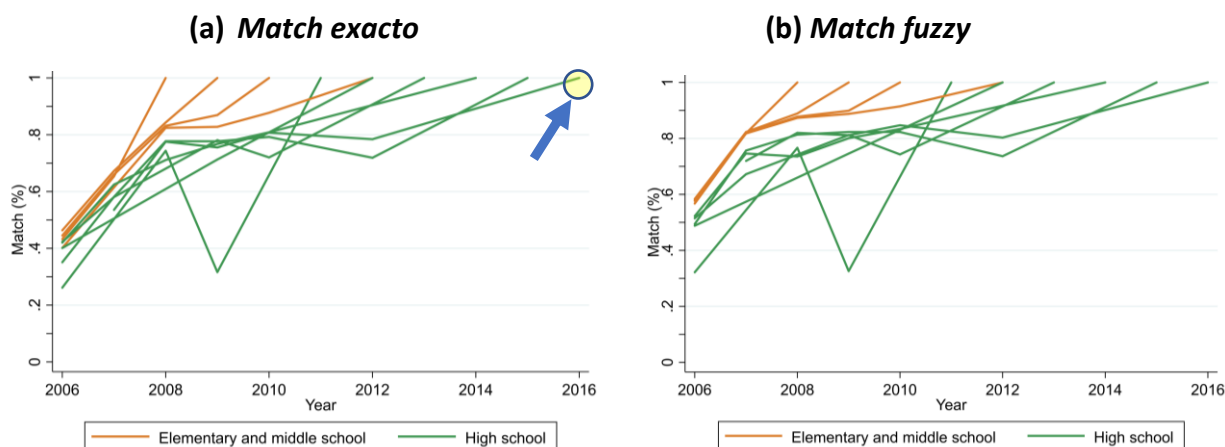
Notas: Usamos los datos cuyo cuarto y quinto carácter del CCT correspondiera a Escuela de Educación Secundaria General (ES), Escuela de Educación Secundaria Técnica (ST) y Telesecundaria (TV).

La Tabla 4 muestra las generaciones en renglones y los números que deberíamos tener para cada una. Sacando un promedio simple² obtenemos que deberíamos tener cerca de 3.08 años de seguimiento por alumno, y en realidad tenemos 2.62 en el *panel_fuzzy*, es decir tenemos un 85% 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.

Como es bien sabido, Enlace es una prueba estandarizada que se hizo desde el 2006 al 2013 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.³

Para dar una primera idea de la calidad de los datos, la Figura 1 muestra el porcentaje de observaciones encontradas en el *Panel_exacto* y en el *Panel_fuzzy*.

Figura 1: Porcentaje de individuos encontrados de forma Retrospectiva



Notas: En el anexo se pueden observar gráficas mostrando el porcentaje de individuos encontrados de forma retrospectiva por estado.

Se explica el panel (a) de la Figura 1 para dar una primera idea de la calidad de los datos. Nótese que tomadas de derecha a izquierda todas las líneas empiezan cuando el valor del eje Y es igual a 1. Esto representa el 100%. Tomemos el punto amarillo señalado con una flecha. En este punto nos

² Para mayor exactitud se podría también hacer un promedio ponderado por el número de alumnos en cada generación.

³ Para este reporte sólo se utilizaron los datos de las pruebas Enlace de 2006 a 2012 para primarias. En los años 2006, 2007 y 2008 no tenemos los datos de primero y segundo de secundaria. En el año 2011 únicamente tenemos los datos de primaria.

enfocamos en la generación que estaba en 3ro de preparatoria en el 2016, y luego buscamos a estos mismos alumnos usando el CURP en años anteriores, desde secundaria hasta tercero de primaria. Podemos observar que de la línea que empieza en el 2016, encontramos a casi el 80% en el 2008. Si nos enfocamos solo en primaria (las líneas naranjas), para la generación que estaba en primaria o secundaria en 2012, encontramos a más del 80% en el 2009. Dado que parte de esto puede deberse a inasistencias el día de la prueba, consideramos que es razonable encontrar este número de estudiantes. En secciones posteriores se mostrarán análisis más detallados sobre la calidad de los datos.

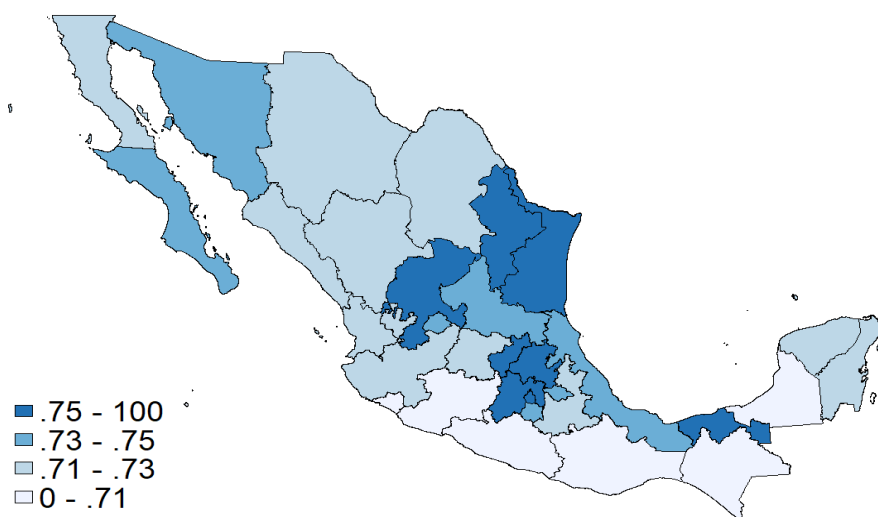
Tabla 4.3. Número de alumnos que participaron en 2006-2012 por Estado

<i>Entidad</i>	<i>2012</i>	<i>2011</i>	<i>2010</i>	<i>2009</i>	<i>2008</i>	<i>2007</i>	<i>2006</i>
<i>Aguascalientes</i>	171,250	104,279	166,736	160,652	115,050	116,655	49,762
<i>Baja California</i>	408,611	250,970	390,050	377,536	276,546	268,122	10,746
<i>Baja California Sur</i>	80,506	49,460	77,203	73,885	47,204	50,650	24,956
<i>Campeche</i>	104,958	64,895	101,578	100,237	67,569	72,130	67,136
<i>Chiapas</i>	362,891	229,657	348,723	636,888	481,462	246,558	111,234
<i>Chihuahua</i>	79,075	45,267	74,638	416,019	303,884	54,120	51,907
<i>Coahuila</i>	557,486	467,740	702,559	335,765	247,374	128,218	2,981
<i>Colima</i>	427,609	262,324	421,926	74,337	52,119	296,910	51,352
<i>Distrito Federal</i>	1,018,002	612,950	1,042,533	1,041,505	724,867	740,619	51,389
<i>Durango</i>	230,393	143,804	218,909	221,226	145,183	163,605	17,911
<i>Guanajuato</i>	799,944	502,639	762,075	745,725	550,547	546,010	106,853
<i>Guerrero</i>	80,825	289,929	441,432	335,440	364,669	355,091	46,684
<i>Hidalgo</i>	371,399	219,477	369,025	357,679	262,456	271,062	68,839
<i>Jalisco</i>	978,287	619,240	936,879	914,105	638,600	656,530	126,575
<i>Michoacán</i>	1,994,410	1,218,751	1,966,434	201,889	3,132	869,303	575
<i>Morelos</i>	153,885	171,972	225,156	191,619	156,906	196,632	60,372
<i>México</i>	226,118	140,241	225,677	1,891,108	1,407,485	163,772	22,802

<i>Nayarit</i>	139,016	84,887	138,690	132,379	93,313	94,216	29,304
<i>Nuevo León</i>	607,695	370,991	569,823	540,933	389,016	380,454	70,828
<i>Oaxaca</i>	7,969	4,214	6,033	64,500	344,282	324,414	
<i>Puebla</i>	838,892	521,121	788,231	670,715	560,371	541,101	52,244
<i>Querétaro</i>	258,567	162,004	255,071	244,902	172,159	175,546	4,934
<i>Quintana Roo</i>	170,678	101,834	161,863	154,284	112,951	115,704	15,294
<i>San Luis Potosí</i>	378,926	236,634	363,110	356,752	250,911	266,155	72,397
<i>Sinaloa</i>	361,701	217,911	349,522	349,852	257,320	266,645	36,731
<i>Sonora</i>	350,792	204,056	340,298	315,993	233,756	230,476	85,367
<i>Tabasco</i>	298,534	183,710	301,185	295,279	207,596	215,205	68,084
<i>Tamaulipas</i>	412,793	259,841	409,099	400,607	288,872	278,878	78,561
<i>Tlaxcala</i>	161,295	96,266	158,846	162,186	114,024	119,665	41,827
<i>Veracruz</i>	1,022,253	632,785	1,010,051	994,410	730,392	728,421	56,723
<i>Yucatán</i>	250,633	157,330	242,380	234,255	170,017	172,113	98,546
<i>Zacatecas</i>	201,774	131,328	206,624	195,020	140,852	143,458	48,579

Notas: El número de alumnos que participaron por año y por entidad se contaron usando el folio, pues el número de alumnos contados por CURP es mucho menor por los CURP faltantes.

Mapa con el *ratio* de Enlace entre niños en edad de asistir en el estado



Ratio Enlace: Tomamos el número de alumnos inscritos por entidad y el número de alumnos por entidad que encontramos en el panel que armamos 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.

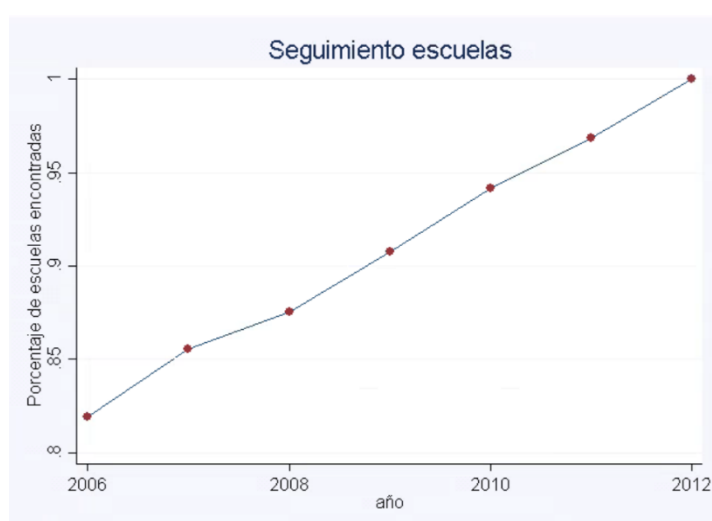
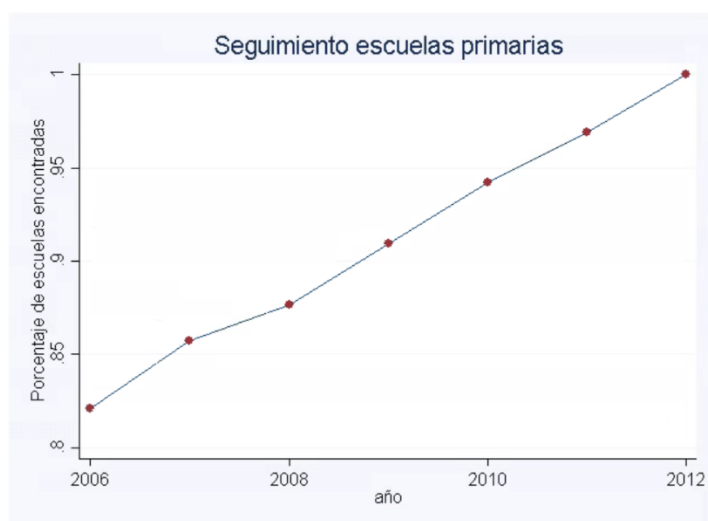
Tabla 4.4. Número de escuelas que participaron en 2006-2012 por Estado

<i>Entidad</i>	<i>2012</i>	<i>2011</i>	<i>2010</i>	<i>2009</i>	<i>2008</i>	<i>2007</i>	<i>2006</i>
<i>Aguascalientes</i>	1,030	745	1,020	1,022	977	993	566
<i>Baja California</i>	2,178	1,676	2,084	2,029	1,974	1,919	454
<i>Baja California Sur</i>	571	419	551	530	490	502	457
<i>Campeche</i>	1,081	785	1,022	1,054	1,024	1,026	1,019
<i>Chiapas</i>	2,355	1,861	2,217	10,084	9,577	2,190	1,109
<i>Chihuahua</i>	643	465	629	3,445	3,310	607	567
<i>Coahuila</i>	7,964	8,265	9,974	2,294	2,210	4,370	2,864
<i>Colima</i>	3,556	2,751	3,469	621	602	3,254	807
<i>Distrito Federal</i>	4,554	3,268	4,611	4,620	4,619	4,622	269
<i>Durango</i>	3,414	2,520	3,384	3,373	3,160	3,132	566
<i>Guanajuato</i>	6,287	4,690	6,194	6,134	6,056	6,079	2,143
<i>Guerrero</i>	1,893	4,282	5,599	4,485	5,928	5,801	1,927
<i>Hidalgo</i>	4,369	3,205	4,352	4,320	4,170	4,237	1,921
<i>Jalisco</i>	7,553	5,873	7,420	7,372	7,126	7,144	3,040
<i>Michoacán</i>	11,274	7,708	11,124	3,627	745	6,678	1,380
<i>Morelos</i>	1,772	3,053	3,584	1,358	1,388	3,104	1,696
<i>México</i>	1,513	1,082	1,470	10,975	10,862	1,393	194
<i>Nayarit</i>	1,708	1,185	1,681	1,651	1,617	1,597	1,290
<i>Nuevo León</i>	3,686	2,765	3,541	3,452	3,345	3,273	797
<i>Oaxaca</i>	1,192	846	1,089	1,841	6,697	6,060	
<i>Puebla</i>	6,663	4,564	6,525	5,536	6,390	6,342	1,371
<i>Querétaro</i>	1,961	1,461	1,904	1,881	1,605	1,799	106

<i>Quintana Roo</i>	1,172	822	1,123	1,061	1,020	1,006	283
<i>San Luis Potosí</i>	4,963	3,377	4,939	4,933	4,839	4,790	1,878
<i>Sinaloa</i>	3,545	2,742	3,465	3,447	3,356	3,328	1,059
<i>Sonora</i>	2,441	1,678	2,414	2,401	2,332	2,250	941
<i>Tabasco</i>	2,726	2,025	2,711	2,664	2,625	2,732	977
<i>Tamaulipas</i>	3,162	2,492	3,090	3,093	3,035	3,000	1,072
<i>Tlaxcala</i>	1,086	785	1,069	1,059	988	1,033	452
<i>Veracruz</i>	12,999	9,773	12,653	12,497	12,087	12,056	1,351
<i>Yucatán</i>	1,951	1,369	1,880	1,826	1,787	1,803	1,785
<i>Zacatecas</i>	3,084	2,005	3,117	3,096	3,042	3,076	2,008

Notas: En algunos años, el número de escuelas es menor porque la prueba no se implementó a todos los niveles, es decir, en los años en los cuales no se aplicó en secundaria, el número de escuelas es menor.

Figura 2: Porcentaje de escuelas encontrados de forma Retrospectiva



Notas: Tomamos a las escuelas que aparecen en 2012 e hicimos un seguimiento de ellas en el pasado: hasta 2006 encontramos al 83%.

Objetivo 2: Calidad del panel y los datos

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

A) Existencia de las observaciones

La Tabla 5 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.

Tabla 5: Generaciones de alumnos, grados escolares y años con Enlace

<i>Años escolares</i>	2012	2011	2010	2009	2008	2007	2006
<i>3ro prim</i>	X	X	X	X	X	X	X
<i>4to prim</i>	X	X	X	X	X	X	X
<i>5to prim</i>	X	X	X	X	X	X	x
<i>6to prim</i>	X	X	X	X	X	x	X
<i>1ro secu</i>	X		X	X			
<i>2do secu</i>	X		X	x			
<i>3ro secu</i>	X		x	X	X	x	X

Lo que se observa aquí es que para algunas generaciones podemos seguir a los alumnos por varios años. Por ejemplo, la generación roja 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.⁴ Cabe notar que, en 2006, 2007 y 2008, segundo y primero de secundaria no presentaron la prueba Enlace y que la base que nos proporcionaron de 2011 solo tiene primaria.

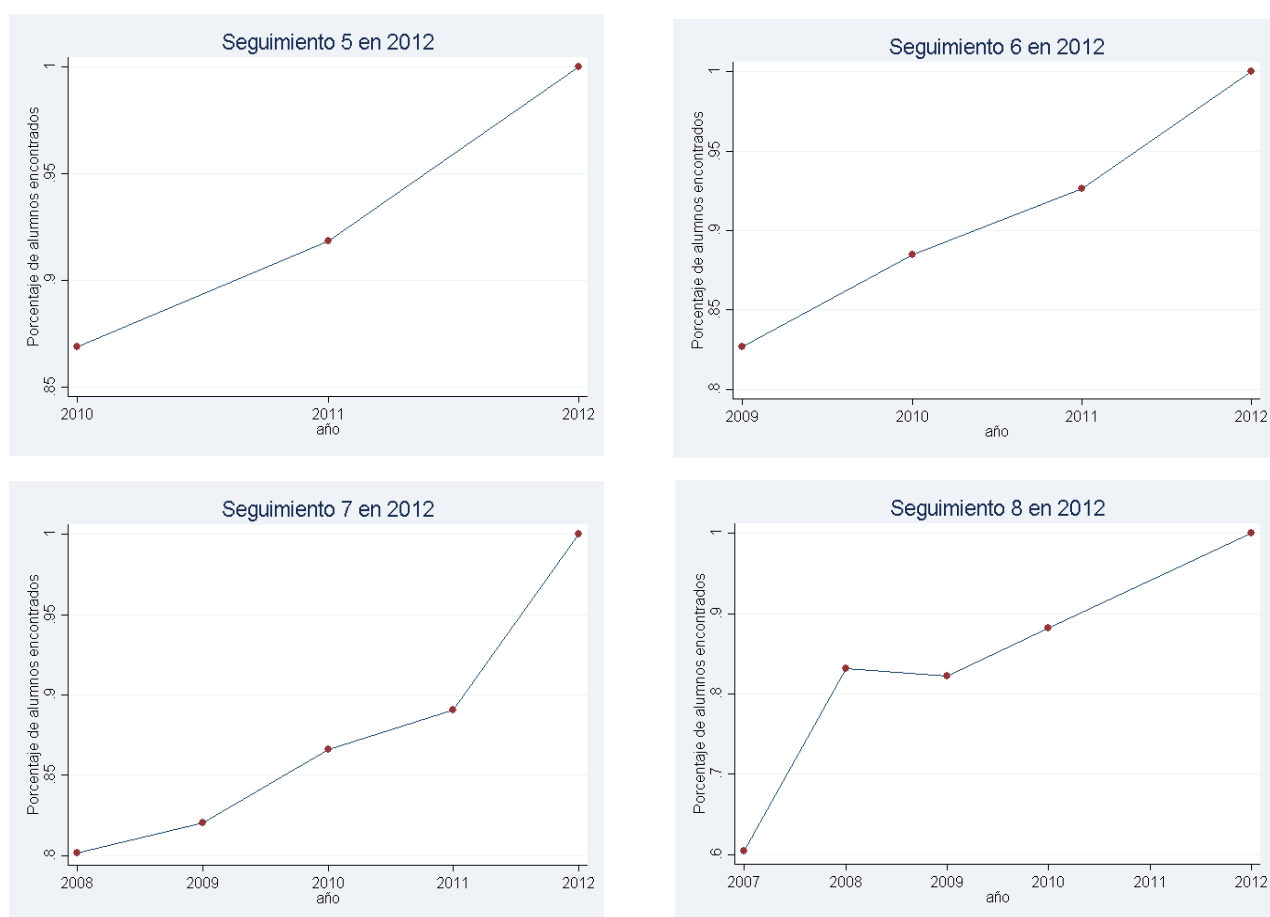
A diferencia de la Figura 1 que busca a los niños de un año en años pasados (incluidos los de Prepa, usando Planea), la Figura 2 hace un seguimiento retrospectivo de una generación en específico sin usar Planea. Esto añade calidad al análisis porque permite identificar si los datos faltantes se dan en

⁴ Como no usamos las pruebas PLANEa son solo 6 las generaciones que podemos seguir en el tiempo por al menos tres años.

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 2a se lee de la siguiente manera. Si el título de la gráfica es “seguimiento 6 en 2012” tomamos a los niños que iban en sexto de primaria en el año 2012 y los buscamos en años anteriores. Es decir, 6 corresponde a sexto de primaria, 7 corresponde a primero de secundaria, etc. Por ejemplo, en 2011 encontramos únicamente al 93% de los niños que iban en sexto de primaria en 2012.

Figura 2a: A cuántos alumnos encontramos retrospectivamente a partir del último año observado.



La Figura 2 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.⁵

La Figura 2a toma las cuatro generaciones para las que observamos toda la primaria y, empezando de 6to de primaria, se pregunta que fracción de personas se encuentran en años escolares anteriores.

⁵ En el panel izquierdo superior, solo observamos quinto cuarto y tercero de primaria, por eso hay menos puntos.

Figura 2b: 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.

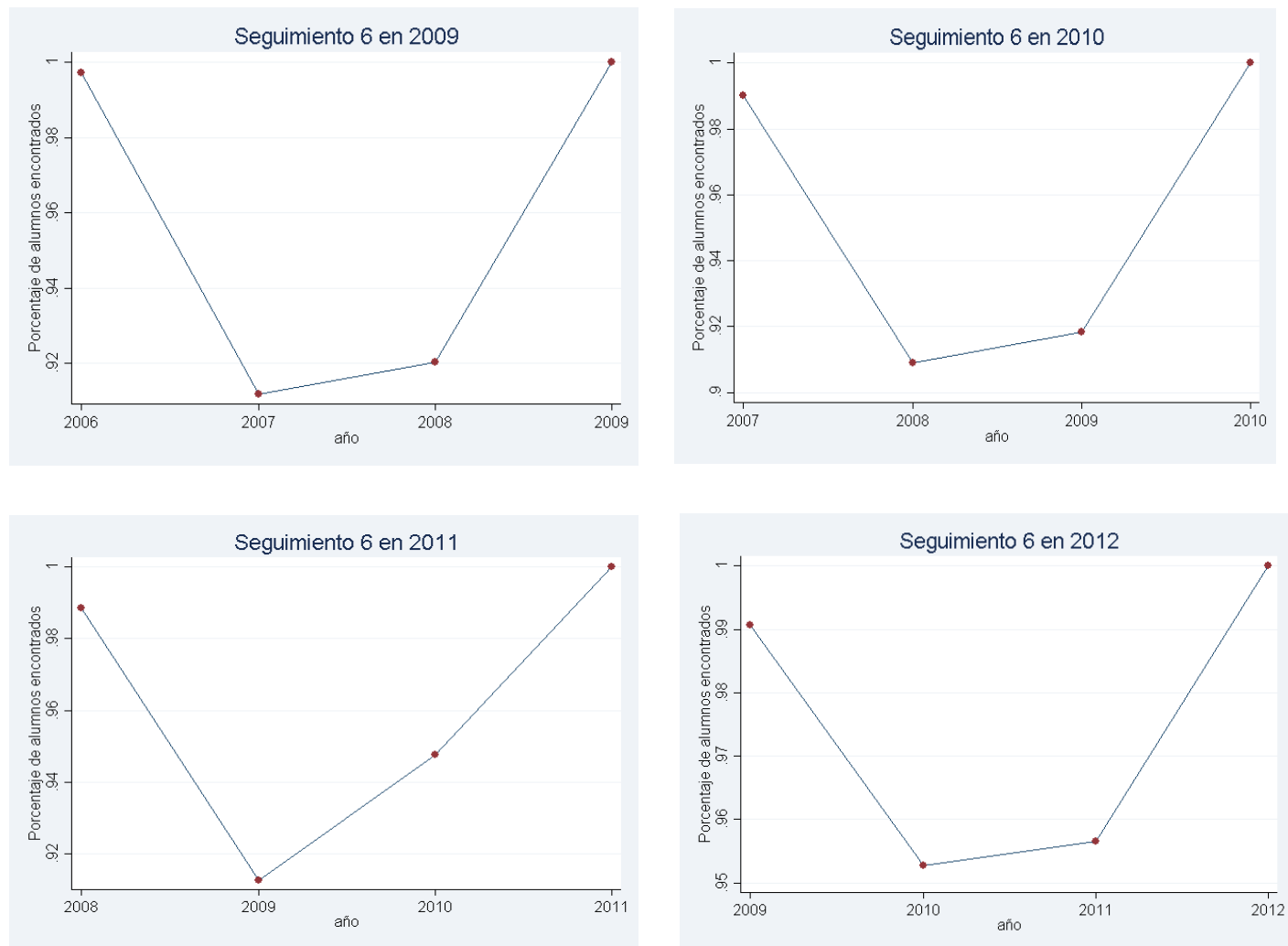
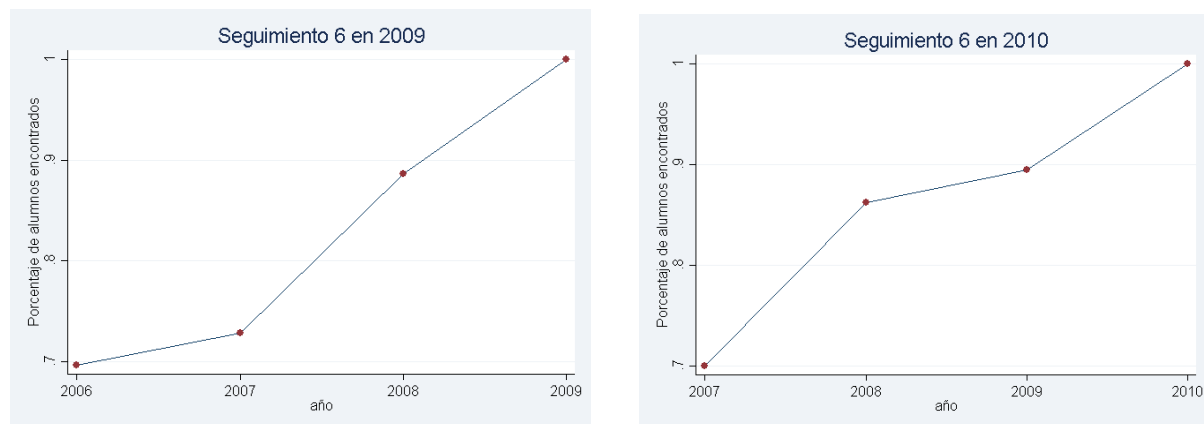


Figura 2c: A cuántos alumnos encontramos retrospectivamente a partir del último año observado condicionando en que tenemos información de la escuela a la que asistieron en todos los años.



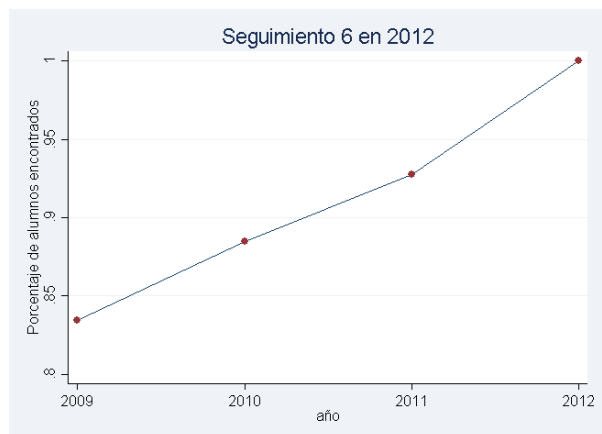
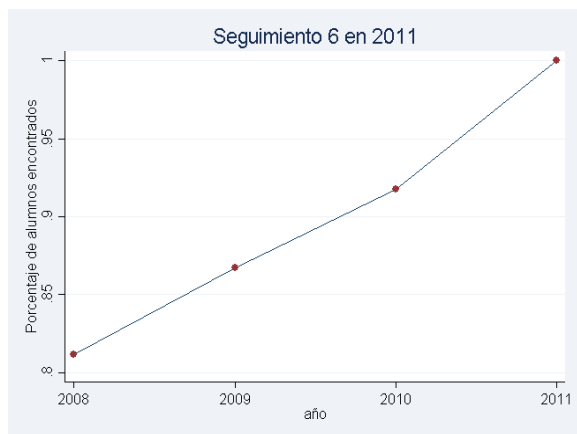
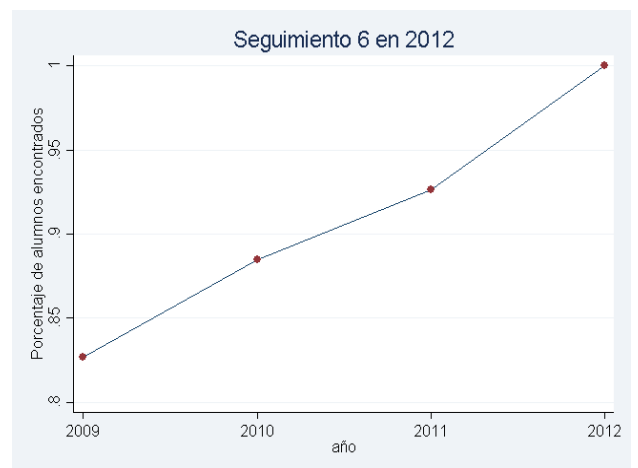
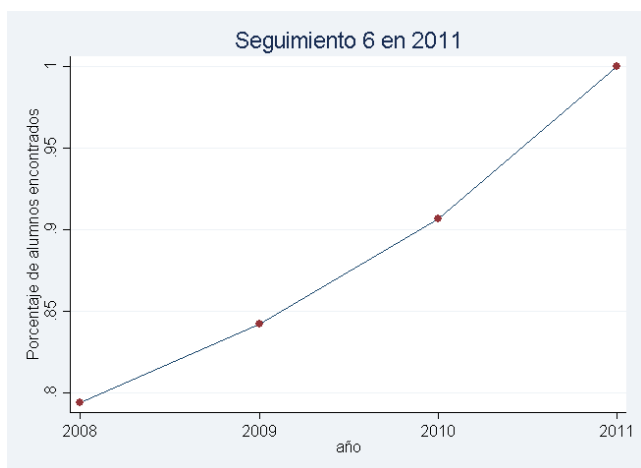
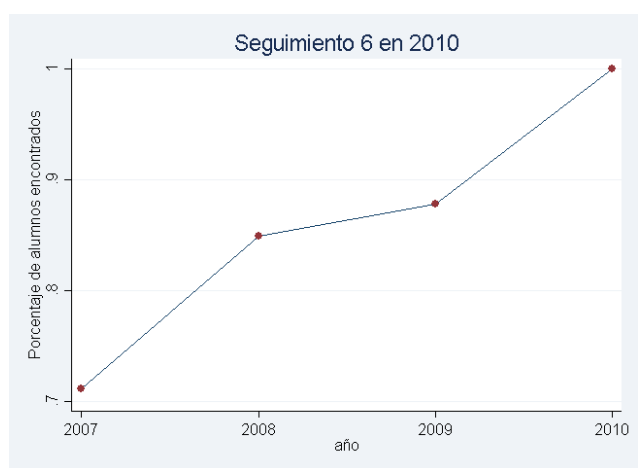
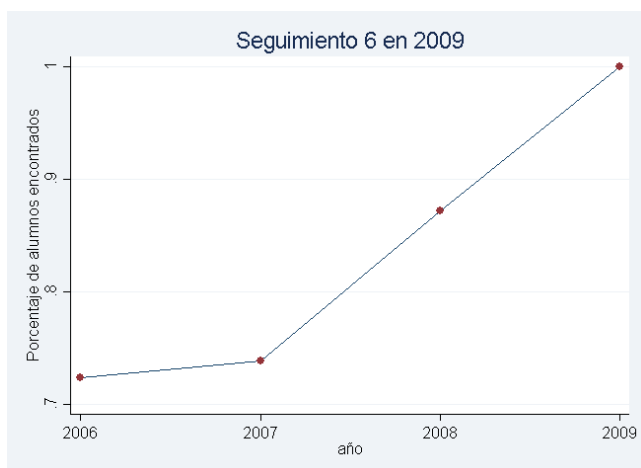


Figura 3: A cuántos alumnos encontramos retrospectivamente desde sexto de primaria empezando en distintos años



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

Las gráficas de la figura 2c muestran una pendiente positiva del año 2011 hasta el 2008, 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 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.

Por último, quisiéramos decir algo sobre las inasistencias, pero para hacer esto debemos tener cuidado en distinguir inasistencias de deserción, o de casos en los que por los periodos muestrales no tenemos información. Definimos una inasistencia como: el alumno no presentó examen en un año, pero al año siguiente sí asistió a la escuela. Si el alumno no presenta un examen un año y tampoco asiste a la escuela en años siguientes podría tratarse de una persona que falta mucho o de un caso de abandono escolar, como no es posible diferenciar entre ambos este tipo de casos no lo tratamos como inasistencia.

Nótese que identificar inasistencias no es trivial. En el último año en el que observamos a la generación, por ejemplo, no podemos diferenciar entre faltas y abandono escolar porque no tenemos información de los siguientes años para saber si el alumno continúa inscrito después de esa falta. Entonces, tenemos que usar años “interiores” a la base: si el alumno presentó el examen en un año, significa que en los años anteriores estaba inscrito en la escuela (de otra forma no hubiera podido llegar a ese grado escolar, pues no puede estar en quinto sin haber pasado por tercero y cuarto). Si en los años anteriores su generación presentó el examen, pero el alumno no lo hizo, entonces lo tratamos como inasistencia.

A continuación, presentamos un ejemplo de cómo se calcularon inasistencias para la generación que llamamos “7” anteriormente en la Tabla 2. Cuando analizamos las faltas nos podemos dar cuenta que existen distintos patrones, hay personas que nunca fueron al examen, que nunca faltaron, que faltaron solo un año, etc. En la tabla 6 un renglón es un patrón; ahí, se muestran los patrones que aparecieron en más de 1% de los alumnos de la generación. Las primeras cinco columnas indican cómo se ve el patrón: “0” significa que no presentó el examen; cualquier número distinto que “0” indica el grado en el que iba esa generación en ese año y que sí presentó el examen. Marcados con rojo están los ceros que consideramos como faltas. Si un cero es blanco no se tomó en cuenta para contabilizar las inasistencias. La parte derecha de la tabla indica las inasistencias totales.

El primer renglón de la Tabla 6 muestra que hay 93,905 alumnos de la generación 7 que aparecen en primero de secundaria en 2012 pero que no aparecen en ningún año del 2008 al 2012. Los cuadros en rojo indican lo que creemos que son “inasistencias” (o errores): sabemos que el niño debió estar en alguna escuela, sin embargo, no lo observamos. El segundo renglón toma a la generación 7, de nuevo, que debía estar en sexto de primaria en 2011. Nótese que, en este renglón, en el 2012, hay un

cero, pero no está en rojo porque como no sabemos si desertó al siguiente año, no podemos llamarlo inasistencia. De esta forma, para la generación 7 calculamos inasistencias como el número de ceros en esta tabla por año, y luego dividimos por el número de alumnos de esa generación que observamos ese año.

Por ejemplo, tomemos el renglón 5. Este tiene un 5 en la columna 2010 y un 6 en la columna 2011. Esto quiere decir que 34,919 niños (se lee de la columna “número”) no presentaron el examen en 2009 y 2008 por lo que los tratamos como faltas y en 2010 estaban en quinto de primaria y sí presentaron el examen. En sexto de primaria en 2011 también presentaron el examen. En 2012 no hicieron la prueba ENLACE, pero no lo contamos como falta porque no sabemos si faltaron o abandonaron la escuela.

Después, se suman las faltas de cada año que hubo en esa generación: 555,528 en 2008. Con esto calculamos el porcentaje respecto al total de alumnos que asistían en esa generación en ese año este se calcula de la siguiente manera:

$$\frac{faltas_t}{faltas_t + total\ que\ presentó\ el\ examen_t}$$

para 2008 sería:

$$\frac{555,528}{555,528 + 1,751,077} = 0.24084 \approx 24\%$$

Tabla 6: Ejemplo de cómo se calcularon inasistencias para generación 7 en 2012

Grado						Numero de estos casos
2008	2009	2010	2011	2012		
0	0	0	0	7		93905
0	0	0	6	0		39272
0	0	0	6	7		33723
0	0	5	0	0		28676
0	0	5	6	0		34919
0	0	5	6	7		60117
0	4	0	0	0		61910
0	4	5	0	0		27255
0	4	5	6	0		57158
0	4	5	6	7		118593
3	0	0	0	0		283932
3	0	0	0	7		25245
3	0	5	6	7		56396
3	4	0	0	0		92715
3	4	0	6	7		31212
3	4	5	0	0		57116
3	4	5	0	7		27984
3	4	5	6	0		138720
3	4	5	6	7		956200

Faltas			
2008	2009	2010	2011
555528	372253	223357	147134


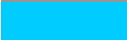





Faltas en porcentaje			
24%	19%	12%	8%

Usamos todo el panel exacto, menos a los alumnos que reprobaron, para generar esta tabla.

En la Tabla 7 hacemos este cálculo para todas las generaciones. Usamos dos símbolos: “.” significa que ese año la generación no presentó el examen. Por ejemplo, la generación que llamamos “5 en 2012” no presentó el examen en 2009 porque en ese año estaban en segundo de primaria y la prueba de Enlace se presentaba a partir de tercero de primaria. El segundo símbolo, “x”, significa que la generación sí presentó el examen, pero no contamos con suficiente información para saber si la ausencia de la calificación es una inasistencia o se trata de un alumno que abandonó sus estudios, por ser el último año que tenemos información de esa generación. Los porcentajes de inasistencias excluyen a alumnos que reprobaron⁶, porque no queda claro a qué generación pertenecen.

Las inasistencias “interiores” rondan entre el 7% y el 19% si excluimos el 2006 y el 2007. Recuerden que estamos llamando “inasistencias” a las lagunas que hay en los datos dentro de los años en los que podemos seguir a un alumno. En particular, nos interesa saber en cuántos casos sucede la intermitencia del alumno en la prueba. Esto podría ser un tema de inasistencias o de ausencia de datos. El hecho de que las inasistencias rondan entre el 7% y el 19% es satisfactorio, ya que según entendemos las inasistencias pueden estar cerca del 15% en la realidad.

Tabla 7: Inasistencias

		2006	2007	2008	2009	2010	2011	2012
	5 en 2012	0.14	0.08	x
	6 en 2012	.	.	.	0.19	0.12	0.06	x
	7 en 2012	.	.	0.24	0.19	0.12	0.08	x
	8 en 2012	.	0.30	0.20	0.11	0.04	.	x
	9 en 2012	0.46	0.28	0.11	0.07	0.02	.	x
	8 en 2010	0.48	0.27	0.12	0.11	x	.	.
	9 en 2010	0.44	0.21	0.08	x	x	.	.

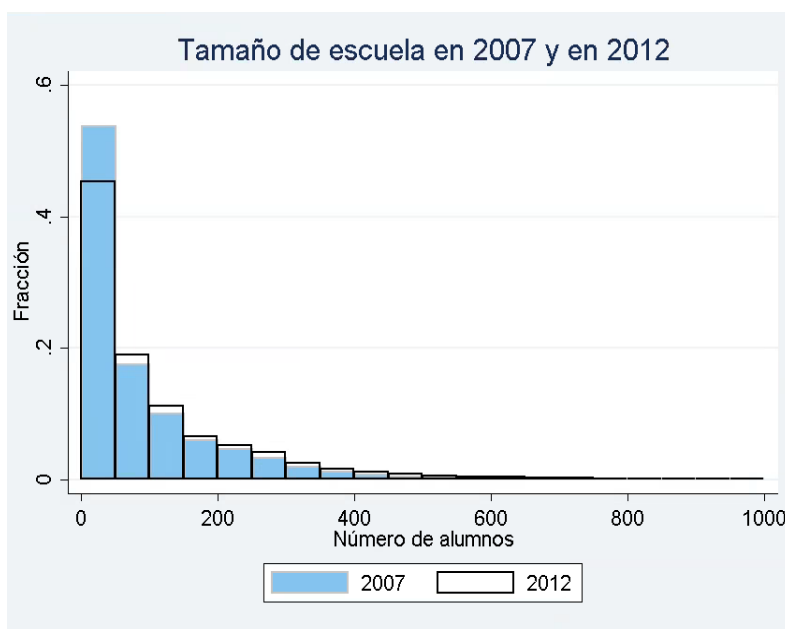
Así pues, notamos que la base de datos del 2006 es de menor calidad debido a las siguientes razones: El número de faltas es mucho mayor a comparación del resto de los años. Los porcentajes disminuyen mucho conforme nos acercamos a los últimos datos observados por generación, esto puede deberse a que, dado nuestro criterio de faltas (si el alumno no presenta el examen en un año pero sí lo presenta en algún año subsecuente entonces lo tratamos como falta) puede haber ausencias tratadas como abandono escolar y no como faltas. Por ejemplo, un alumno que falte al examen en 2010, 2011 y 2012 no lo contamos como inasistencia porque no tenemos manera de distinguirlo de un caso de abandono escolar. Cuanto más cercano está el año en cuestión al 2012 más probable es que tratemos un caso de falta como abandono escolar porque hay menos años posteriores donde el alumno pudo haber presentado la prueba.

⁶ Es decir, aparecen en el mismo grado en dos años calendarios consecutivos.

B) Tamaño de matrícula

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.

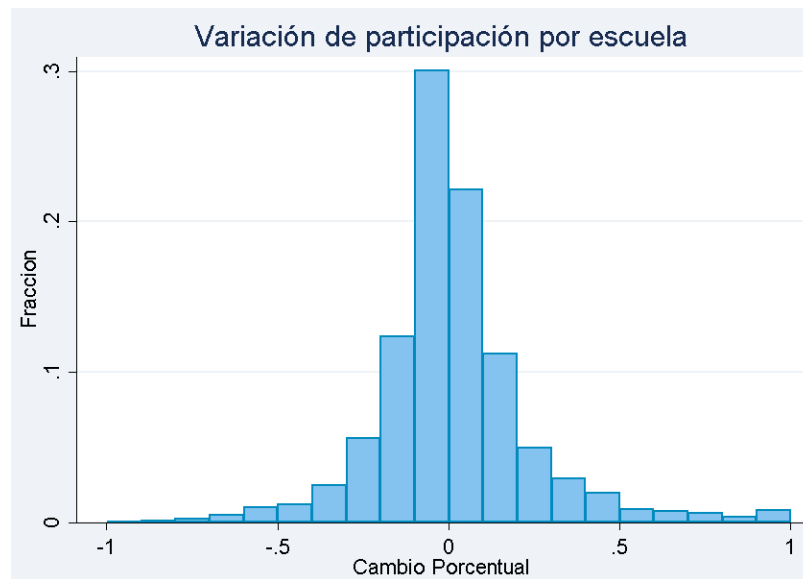
Figura 4: Histograma del tamaño de las escuelas en promedio en 2007 y 2012



Notas: 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 5 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 5: Variación en asistencias: la misma escuela a través del tiempo



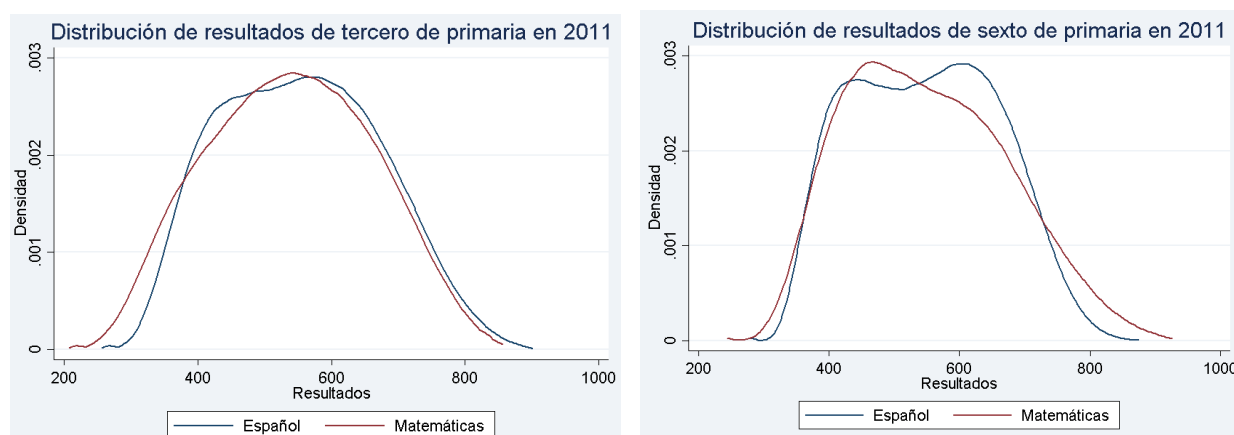
Notas: La variación se calculó como el cambio porcentual del tamaño de la escuela entre años consecutivos. Por cada año, para cada escuela se calculó el número de alumnos que participaron en la prueba y se comparó con el año siguiente para encontrar el cambio porcentual.

c) 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 6: Distribución de calificaciones en Matemáticas y Español para para 3ro de primaria y 6to de primaria, y para 3ro de secundaria, año 2011.

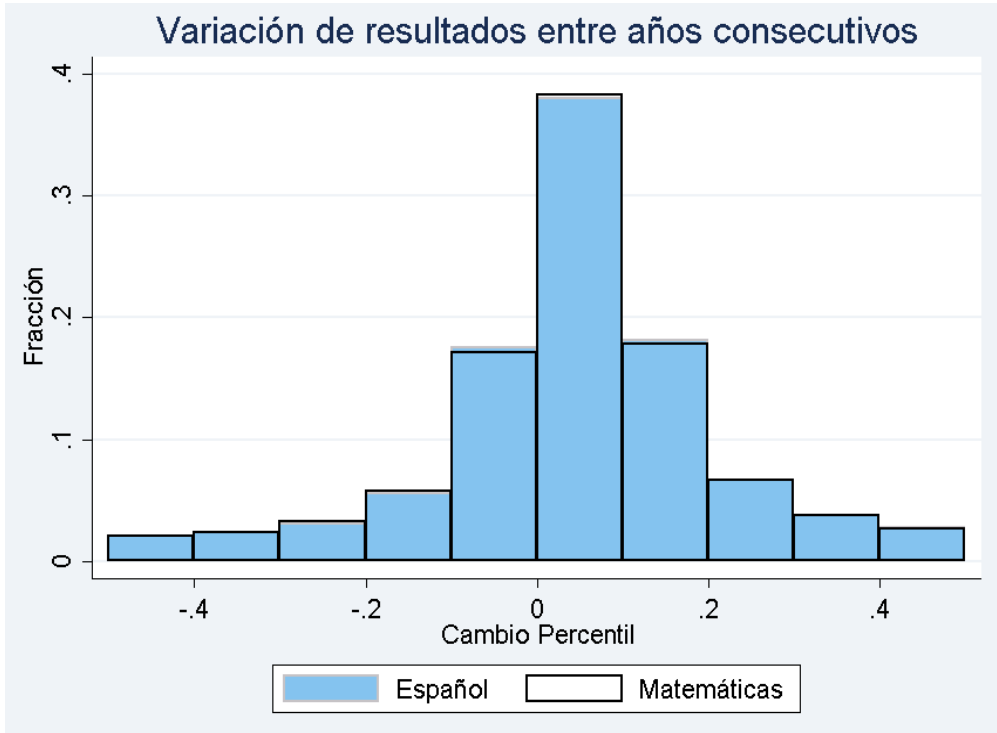


Notas: Se muestra a la distribución del 2011 como caracterización de los otros años. Es decir, los resultados se comportan de forma muy parecida en los otros años y los otros grados

Estabilidad en el tiempo por escuela

La Figura 6 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.

Figura 7: Variación de calificaciones entre años, por escuela



Notas: El cambio percentil se refiere al cambio en la posición del alumno relativa a los otros alumnos de su generación. Se generó un ranking de Español y otro de Matemáticas y utilizando el ranking se calculó el percentil al cual pertenece el alumno. Acto seguido, se calculó el cambio en percentiles para años consecutivos.

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: $C_{it} = \alpha + \beta C_{i,t-1} + \varepsilon_{it}$, donde C_{it} es la calificación del alumno “i” en el año “t”.

Tabla 8: Persistencia a nivel persona (regresión)

VARIABLES	(1) p_mat_std	(2) p_esp_std
L.p_mat_std	0.664*** (0.000)	
L.p_esp_std		0.668*** (0.000)
Constant	-0.000* (0.000)	0.003*** (0.000)
Observations	23,884,096	23,860,344
R-squared	0.434	0.440

Robust standard errors in parentheses

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

D) 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) cambiaran el 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.

Cierres

Tabla 9: Escuelas cerradas por años para primarias y secundarias (definiciones múltiples)⁷

año	total	número de escuelas que dejan de aparecer entre un año y otro	porcentaje	25% de alumnos encontrados al siguiente año	porcentaje	40% de alumnos encontrados al siguiente año	porcentaje
2006	111253	1146	1.03%	1014	0.91%	926	0.83%
2007	119893	1214	1.01%	863	0.72%	724	0.60%
2008	119957	2664	2.22%	1624	1.35%	1492	1.24%
2009	117732	2410	2.05%	2098	1.78%	1801	1.53%
2010	119843	5030	4.20%	2972	2.48%	1686	1.41%

⁷ No se añadió 2011 porque solo se tiene para primarias y esta tabla incluye secundarias.

Tabla 10: Escuelas cerradas por años para primarias públicas (definiciones múltiples)⁸

año	número de escuelas que dejan de aparecer entre un año y otro	porcentaje	25% de alumnos encontrados al siguiente año	porcentaje	40% de alumnos encontrados al siguiente año	porcentaje	total
2006	963	1.21%	862	1.08%	795	1.00%	79475
2007	969	1.14%	719	0.85%	607	0.71%	84949
2008	2243	2.67%	1374	1.64%	1277	1.52%	83995
2009	1391	1.72%	1256	1.55%	1199	1.48%	80779
2010	398	0.49%	311	0.38%	249	0.30%	81844
2011	4413	5.34%	2208	2.67%	1825	2.21%	82585

Tabla 11: Escuelas cerradas por años para primarias privadas (definiciones múltiples)⁹

año	número de escuelas que dejan de aparecer al siguiente año	porcentaje	25% de alumnos encontrados al siguiente año	porcentaje	40% de alumnos encontrados al siguiente año	porcentaje	total
2006	106	1.65%	88	1.37%	67	1.05%	6410
2007	122	1.74%	87	1.24%	64	0.91%	6995
2008	151	2.15%	134	1.90%	120	1.71%	7035
2009	180	2.40%	131	1.75%	116	1.55%	7486
2010	139	1.82%	117	1.53%	91	1.19%	7658
2011	252	3.19%	165	2.09%	148	1.88%	7891

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.

⁸ Si se añadió 2011 porque esta tabla solo incluye primarias y si contamos con esa información en 2011.

⁹ Si se añadió 2011 porque esta tabla solo incluye primarias y si contamos con esa información en 2011.

- **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 12: Escuelas que se incorporan a la base

año	total	abre		25%		40%	
2007	119893	487	0.41%	313	0.26%	277	0.23%
2008	119957	3882	3.24%	2075	1.73%	1755	1.46%
2009	117732	1512	1.28%	1198	1.02%	1118	0.95%
2010	119843	1915	1.60%	1710	1.43%	1608	1.34%
2012	114268	10358	9.06%	8020	7.02%	6614	5.79%

Tabla 12: Escuelas que se incorporan a la base, primarias publicas

año	número de escuelas que empiezan de aparecer en un año	porcentaje	25% de alumnos encontrados al año anterior	porcentaje	40% de alumnos encontrados al año anterior	porcentaje	total
2007	5011	6.31%	2853	3.59%	2367	2.98%	79475
2008	1128	1.33%	994	1.17%	887	1.04%	84949
2009	1773	2.11%	1666	1.98%	1611	1.92%	83995
2010	1436	1.78%	1299	1.61%	1227	1.52%	80779
2011	483	0.59%	331	0.40%	251	0.31%	81844
2012	1281	1.55%	1204	1.46%	1169	1.42%	82585

Tabla 13: Escuelas que se incorporan a la base, primarias privadas

año	número de escuelas que empiezan de aparecer en un año	porcentaje	25% de alumnos encontrados al año anterior	porcentaje	40% de alumnos encontrados al año anterior	porcentaje	total
2007	396	6.18%	140	2.18%	102	1.59%	6410
2008	235	3.36%	150	2.14%	106	1.52%	6995
2009	238	3.38%	134	1.90%	79	1.12%	7035
2010	278	3.71%	159	2.12%	106	1.42%	7486
2011	216	2.82%	143	1.87%	97	1.27%	7658
2012	297	3.76%	210	2.66%	159	2.01%	7891

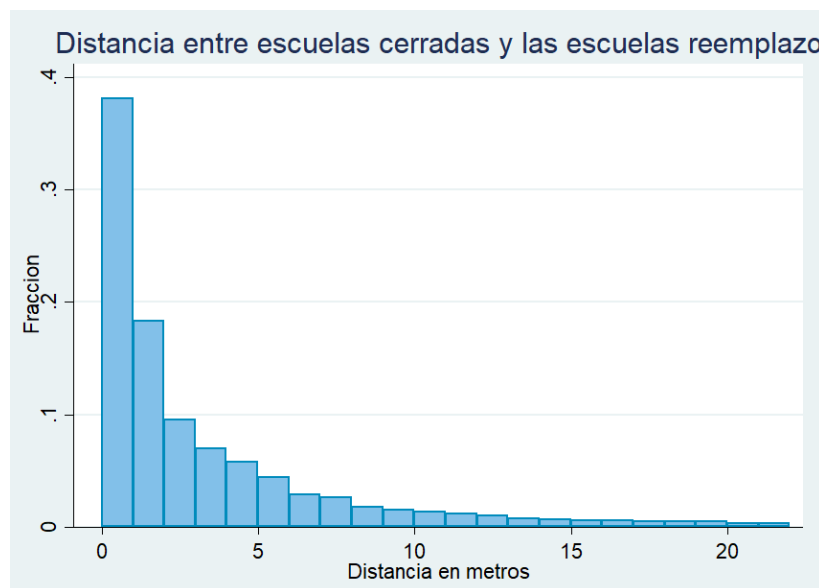
Distancias

Aunque no es parte de la información proporcionada por el Banco Mundial, y no está en el código de Stata que reportamos, 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 8 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 8: Distancias entre escuelas cerradas y las escuelas reemplazo



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

Porcentaje de individuos encontrados de forma Retrospectiva por entidad federativa

Nota: Los números representan las entidades federativas por orden alfabético, siendo 01 Aguascalientes y 32 Zacatecas.

