XVIII JAECAS Jornadas de Estadística de las Comunidades Autónomas
Cáceres del 20 al 23 de Octubre de 2010

Base de Datos Longitudinal de Población de Andalucía (BDLPA): Modelo de datos y sistema de gestión

Viciana Fernández, Francisco
franciscoj.viciana@juntadeandalucia.es
Montañés Cobo, Víctor
victor.montanes@juntadeandalucia.es
Cánovas Balboa, Mª Rosa
rosa.canovas@juntadeandalucia.es
Poza Cruz, Eva
evav.poza.ext@juntadeandalucia.es
Instituto de Estadística de Andalucía

Resumen

La "BDLPA" integra la información de sistemas hasta ahora independientes. Aquí se describe el modelo de datos usado para almacenar stocks y variaciones padronales, junto con los boletines estadísticos del MNP.

El núcleo del sistema es una clave de persona (IDP) asignada a cada una de los individuos que protagonizan los eventos notificados. El subsistema padronal se organiza en "Altas" y "Bajas" conectadas entre sí mediante relaciones diacrónicas y sincrónicas espacialmente coherentes. Otra tabla de observaciones recoge estados de las personas en una fecha concreta. La información del MNP gira alrededor de una tabla de personas que aparecen en boletines. Los dos sistemas se conectan a través de una tabla auxiliar de "versiones de persona", fundamental para la asignación del IDP.

La información territorial asociada a eventos se almacena en varias tablas relacionadas mediante claves "proxi" de portal y vivienda. La asignación temporal y espacial de las residencias conectan a los actores en un mismo hogar.

La gestión de este sistema precisa: carga de información base, identificación y asignación del IDP, control de duplicados y falsas asignaciones, depuración de secuencias y generación de tablas con itinerarios y episodios susceptibles de explotación estadística.

Palabras clave: *estudios longitudinales, itinerarios vitales, censos virtuales, bases de datos socio-demográficas*

1.Origen de la Base de Datos Longitudinal de Población de Andalucía (BDLPA)

La BDLPA es el sistema que utiliza la Actividad del Plan Estadístico de Andalucía denomina "Registro de Población de Andalucía" (RPA). En el plan se especifica que este proyecto sera un "marco integrado para las estadísticas de población y territorio" y que servirá para generar "estadísticas de itinerarios vitales, así como para potenciar las estadísticas de pequeña área".

La idea rectora detras de estos objetivos es aprovechar estadísticamente la ingente cantidad de información que el nuevo sistema de coordinación padronal comienza a proporcionar, con el objetivo de hacer girar el conjunto de las estadísticas demográficas clásicas entorno a la información individual que el nuevo sistema empieza a permitir.

El núcleo del nuevo sistema es el aprovechamiento del flujo de información existente entre los ayuntamientos y el INE, y de otros registros administrativos como los Registros Civiles. El INE envía al Instituto de Estadística de Andalucía (IEA) las variaciones municipales acumuladas en un único fichero semestral. Este fichero de flujos semestrales incluye nacimientos, defunciones, cambios de residencia e incidencias de gestión que se han incorporado en forma de altas, bajas o modificaciones en dicho periodo. Acumulando esta información se ha obtenido un fichero con la información del sistema desde el año 1996. La problemática asociada al manejo de la información de flujos enviada por el INE se detalló en una comunicación en las XV JECAS de Palma de Mallorca[1].

El objetivo de obtener estadísticas sobre trayectorias vitales precisa incluir la perspectiva longitudinal o histórica en el sistema, el cual no sólo debe contener la información sobre la situación actual de los residentes, sino que precisa conservar la información de las modificaciones producidas en la residencia y en otras características demográficas de los mismos. Para ello se desea diseñar un sistema de información demográfico capacitado para generar biografías vitales individuales a partir de la información ya disponible en el sistema estadístico, pero que hasta ahora se han procesado con este objetivo.

2.Entradas de información en la BDLPA

Las fuentes estadísticas que en la actualidad se esta manejando para reconstruir las biografías vitales de los andaluces son las siguiente:

1.La información de última renovación padronal de 1996.

2.Los eventos recogidos en el circuito del Movimiento Natural de Población: partos, defunciones y matrimonios.

3.Las variaciones (flujos padronales) ocurridos en los padrones andaluces desde el nacimiento del sistema.

4.La información socio-demográfica recogida de la población residente en el censo de 2001

El núcleo del sistema es sin duda, la información proveniente de sistema de coordinación de los padrón. La información que este proceso suministra se recibe de acuerdo a la estructura especificas para los intercambio INE-Municipio que se resumen en la Ilustración 1. La característica fundamental de este fichero es contiene solo la información que envía un municipio concreto. Por ejemplo una variación residencias viene en dos registros distintos cada uno para el municipio de Alta y otro en de Baja.

[1] F. Viciana, V. Montañés, J.M. Alba, D. Martínez. Actualización del Registro Estadístico de Población de Andalucía (REPA) a partir de los ficheros de variaciones del Padrón Continuo. Primer año de experiencia. Comunicación de las XV JECAS en Palma de Mallorca.

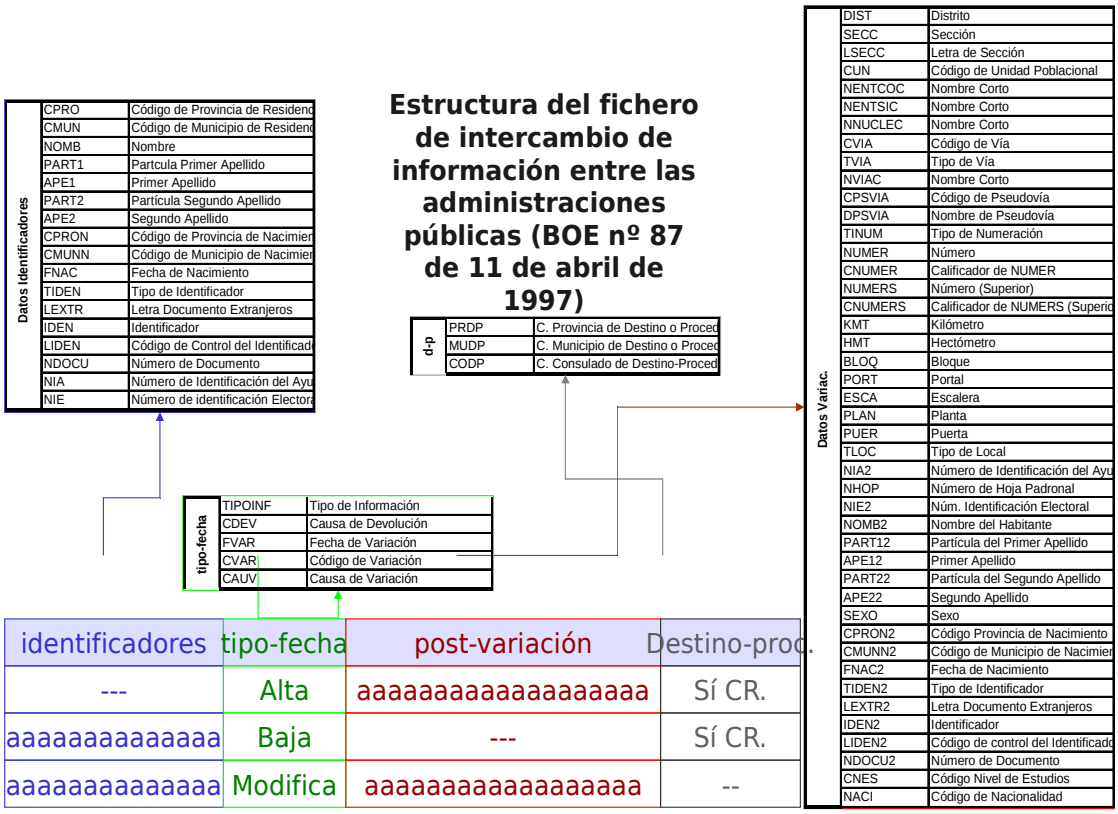


Ilustración 1: Estructura del fichero de intercambio con la información semestral de variaciones suministrado por el INE

3.Esquema conceptual y definiciones.

La idea de para reconstruir biografías vitales y residenciales que puedan ser explotadas estadísticamente, para ello queremos tener un sistema de información capaz de reconstruir las líneas de vida individuales de un supuesto diagrama de Lexis del conjunto de la población andaluza, similar al que se muestra en la Ilustración 2. Disponer de este sistema permitirá la ampliación del tipo estudios a nuevos actividades hasta ahora vedados a la investigación estadística en nuestro medio como son el estudio de las duraciones desde un evento anterior, las interrelaciones entre tipos de eventos, nuevo eventos como emancipaciones, la relaciones de los comportamientos demográficos con las estructuras familiares, inclusión de nuevas variables en el análisis no incluidas en los actuales boletines estadísticos, etc...

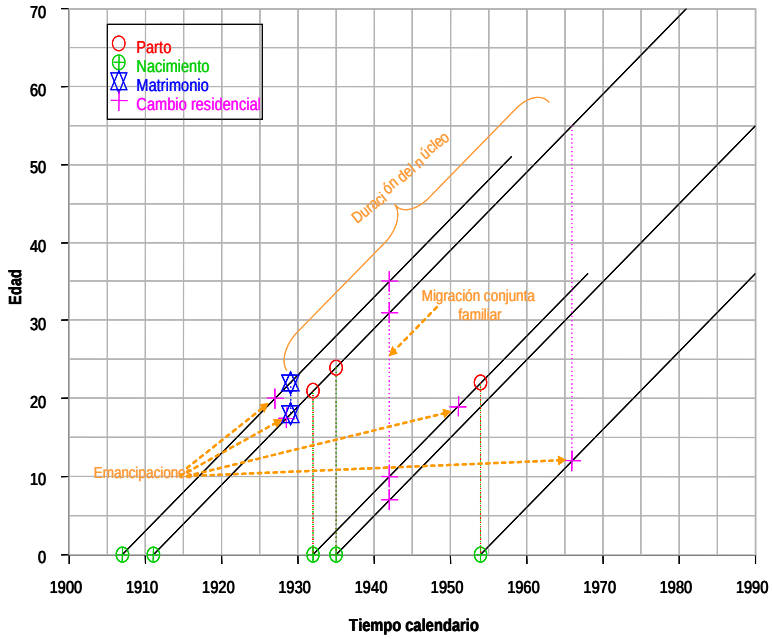


Ilustración 2: Organización de las fuentes de información demográfica sobre las lineas de vida de un diagrama de Lexis

En principio es posible plantear varios modelos distintos para un sistema de este tipo, en una primer nivel es posible platear un sistema basado en una sucesión de observaciones como las que se obtendrían en una encuesta tipo panel, o en los censos repetidos cada diez años. o bien pensar en un sistema basados en vigilancia permanentes con información del momento exacto en que ocurre una transición (Ilustración 3).

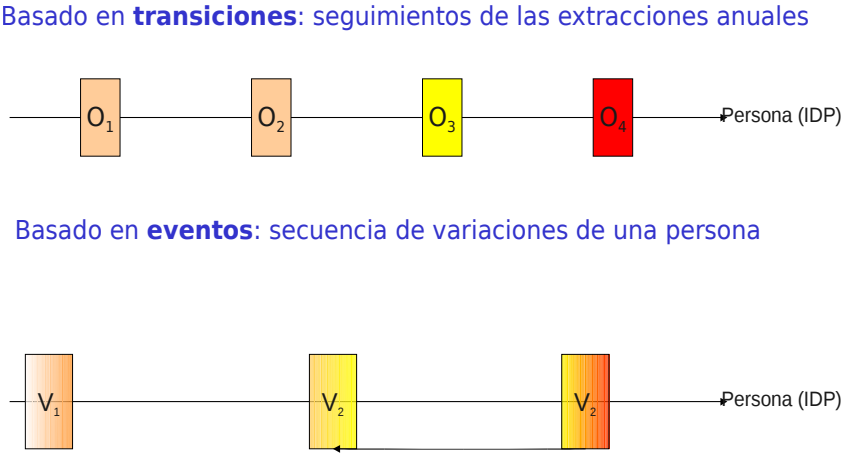
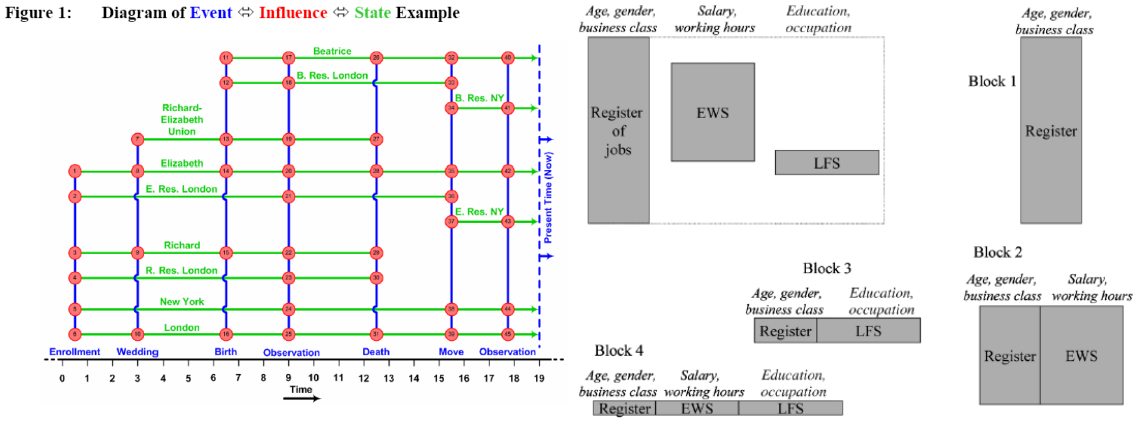


Ilustración 3: Diseño basado en observaciones o basado en transiciones

Parece que un sistema basada en la notificación exacta de los eventos o flujos de variación es preferible, aunque sistema basado en observaciones periódicas, puede resultar más manejable. La cuestión de la manejabilidad, unida a la cuestión de la disponibilidad de información puede hacer atractiva esta opción, de hecho este modelo fue el primero que ensayamos en el IEA a partir de extracciones anuales padronales que nos llegaban del padrón, sin embargo pronto nos dimos cuenta de que la supuesta mayor sencillez de gestión de sistema basado en observaciones era una falsa suposición, al menos para el sistema padronal en España.



A general temporal data model and the structured population event history register.

Demographic Research: Volume 15, Article 7

Ejemplo de enlace de registros y encuestas usadas para obtener la estructura de los ingresos: Sistema de Estadísticas Sociales de Holanda (Social Statistical Database)

Ilustración 4: Dos alternativas diseño centrado en objetivos prioritarios: dinámica o estructura

Otra cuestión importante que se plantea en las primera fases de desarrollo es: ¿Qué objetivo priorizar?. Podemos desear obtener un sistema continuo de seguimiento individual preparado para suministrar información de tipo biográfico. O bien queremos priorizar un sistema orientado a obtener extracciones transversales en determinadas fechas concretas del estado general del conjunto de la población (Censo Virtual). En principio el segundo requisito en más laxo, y por lo tanto sus costes de mantenimiento, parecen menores. Sin embargo tras nuestra experiencia pensamos que incluso para obtener un censo virtual, puede ser preferible pasar previamente por la creación de un sistema de tipo biográfico longitudinal (Ilustración 4)

Con objeto de mostrar algunos de las cuestiones que hemos tenido que ir perfilando durante el desarrollo de sistema, a continuación definiremos algunos conceptos que han resultado especialmente útiles durante el diseño y desarrollo del sistema.

Denominamos "***secuencia registral***" el conjunto de "*apuntes*" en los distintos sistemas que en un momento dado tenemos sobre una determinada persona, identificada en nuestro sistema por tener asignado un mismo IDP. En general los diversos "*apuntes*" o registros que el sistema tiene sobre los individuos se puede clasificar en dos tipos: "*variaciones*" y "*observaciones*".

Denominamos "***variaciones***" a aquellos apuntes que informan de una modificación en algún "*estado*" o característica del individuo y que ha ocurrido en una determinada "*fecha de variación*". Desde un punto de vista teórico puede se conveniente distinguir entre "*fecha real de variación*" y "*fecha de notificación de la variación*" que es la fecha en la cual se declara dicha variación. Piénsese en la altas residenciales, siempre es igual o superior la "*notificación*" al la "*fecha real de variación*". Los diversos "apuntes" en el sistema pueden tener asociada otra marca temporal la "*fecha del apunte"* que es cuando en el sistema se ha almacenado la información. Estas distinciones pueden tener bastante interés en el proceso de depuración e imputación de "itinerarios"

En general los registros del los ficheros de intercambio INE-Municipio que se usan para la coordinación padronal y cuyo clasificación se muestra en la Ilustración 5 tienen, con solo algunos reparos, la consideración de variaciones tal y como las estamos definiendo aquí.

CVAR	CAUV	Descripción
A	OM	Alta por Omisión
	CR	Alta por Cambio de Residencia
	NA	Alta por Nacimiento
B	DE	Baja por Defunción
	II	Baja por inclusión Indebida
	DU	Baja por Duplicado
	CR	Baja por Cambio de Residencia
	BC	Baja por Caducidad ENCSARP
M	PE	Modificación Datos Personales
	RD	Rectificación de Datos por modificaciones Territoriales sin intervención del habitante.
	CD	Modificación Cambio de Domicilio
	RN	Renovación inscripción ENCSARP

Ilustración 5: Tipo de variaciones definidas en sistemas de coordinación padronal

Denominamos "***observaciones***" a aquellos apuntes en el sistema que nos informan que en determinada fecha ("*fecha de la observación*") las características o "estados" que definían a un individuo concreto eran tales. Las observaciones no suelen informan del momento en que se han producido la entrada en dicho "estado". Piense se por ejemplo en un censo, las información recogida en dicha operación tiene la característica de una "observación". O piénsese en el arranque del sistema del padrón continuo con la última renovación padronal de 1996, cuyo arranque son el conjunto de observaciones de la operación de campo de la última renovación padronal de 1996.

Una "secuencia registral" es "***coherente***" si entre los diversos apuntes que la componen ordenado por fecha de variación y observación no hay inconsistencias. Más específicamente si supera un conjunto bien definido de *"reglas de consistencia*". Este conjunto de reglas fundamental en el proceso de depuración e imputación puede ser variable (más o menos estricto) según la extensión del sistema a que se aplique y también del objetivo de la investigación. Por ejemplo una secuencia, del subsistema de variaciones padronales puede ser coherente, pero tornase incoherente cuando se le agregan los eventos provenientes del MNP, igualmente una secuencia puede ser coherente para un determinado nivel de detalle (por ejemplo desagregación municipal) pero ser incoherente a una mayor profundidad de detalle (digamos a nivel de vivienda).

A modo de ejemplo en la Ilustración 6 se muestra una "*secuencia coherente*" de variaciones padronales, para un conjunto sencillo de reglas de consistencia temporal-espacial de "episodios" (no superposición de episodios, un máximo de un episodio abierto y continuidad coherente entre episodio). La biografía registral de este caso identificado en el sistema con el IDP=731,170, se resumen en su nacimiento en 1996 en una vivienda del municipio de Almería (04013), posteriormente en 1999 se traslado a una vivienda del municipio 11004, donde residió hasta 2002, cuando se volvió a trasladar al municipio 04052 en la vivienda 51493 (IDH). En 2005 se cambio de vivienda en el mismo municipio y en 2006 se volvió a cambiar de vivienda y municipio, donde acaba de momento su bibliografía registral.

Tabla de variaciones o Itinerario registrar de IDP = 731,170

Cuadro 2: Variaciones registradas de individuo 731170

	IDV	SSEQ	CV2	FVAR2	PM2	DP2	cola
1	731170	1	ANA0	9625	04013:51430	0	
2	731171	1	BCR1	10591	04013:51430	11004	102
3	731172	1	ACR1	10591	11004:51492	4013	
4	731173	1	BCR1	11787	11004:51492	4052	
5	731174	1	ACR1	11787	04052:51493	11004	
6	731175	1	MCD0	12625	04052:51494	0	
7	731176	1	BCR1	13315	04052:51494	4101	
8	731177	1	ACR1	13315	04101:51495	4052	

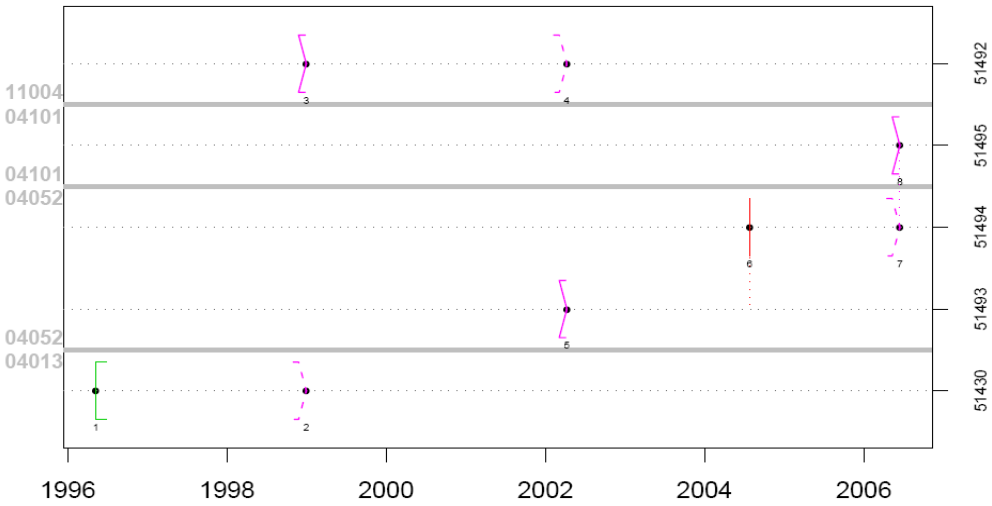


Ilustración 6: Secuencia de variaciones de un IDP, coherentemente encadenadas

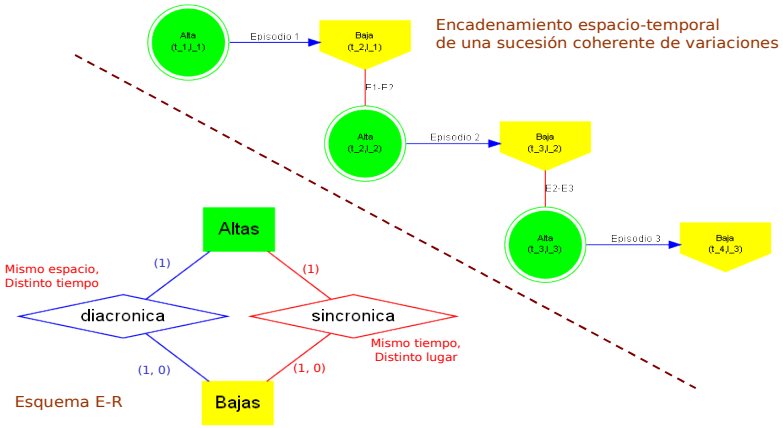


Ilustración 7: Derivación de los episodios desde una sucesión de pares variaciones

Tanto desde un punto de vista teórico, como en los aspectos prácticos de notificación y registro, es conveniente conceptualizar la "*variaciones*" como "***transiciones***", esto es como cambio de estado dentro de un "*espacio de estados*" discreto. Por ejemplo una defunción es una transición del estado vivo a muerto, un "cambio de domicilio" es el cambio de estado residente en la vivienda 1 a residente en la vivienda 2. Desde esta óptica una transición es la combinación de dos sucesos: uno de "baja" en el estado precente y otro de "alta" subsecuentes que ocurren "*sincronicamente*" en la misma "*fecha de variación*". Es posible que dada las fronteras del sistema, no sea posible observar los dos pares de "sucesos", por ejemplo en Andalucia una variación residencial de Madrid a Sevilla se observara exclusivamente como un evento de Alta, al contrario una variación residencial de Sevilla a Madrid aparecerá en el sistema con solo la notificación del alta. Un nacimiento aparece solo como una entrada en Andalucia, sin baja de lugar no especificado. Por ello desde un punto de vista registral "***transición***" podemos definirla como un teórico par de sucesos Baja/Alta entre dos estados distintos, en una misma fecha. Puede que el par de sucesos se notifiquen al sistema en forma de dos registros de variación (como el caso de alta y baja residencial), de uno solo registro con la información de los dos estados (como el caso de las modificaciones del sistema padronal) o como un solo suceso explicito y otro implícito (un nacimiento, una defunción o una entrada o

salida de Andalucía).

Conectado con las "*transiciones*" tenemos el concepto de "***episodio***", que se define como un periodo con duración determinada, delimitado entre un par de sucesos "Alta/baja" que ocurren en distinto tiempo (fecha de variación del alta igual o menor que la de la baja), pero en un mismo estado (por ejemplo en la misma residencia).

La cuestión de las relaciones entre altas y bajas padronales (transiciones y episodios) se entiende mejor si lo mostramos en un esquema donde incluimos el tiempo en horizontal y las transiciones entre estados en vertical. En este caso se representaría una biografía registral coherente de una persona que ha vivido en tres viviendas (h1,h2,h3) y que las variaciones residenciales se han producido en t2,t3. En t1 entro en el sistema digamos por nacimiento y en t4 lo abandono (digamos por defunción)

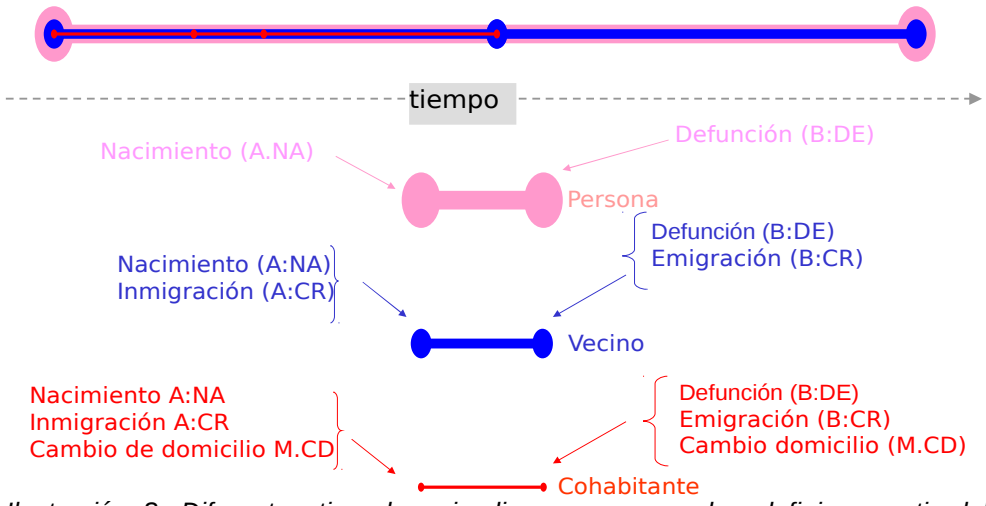


Ilustración 8: Diferentes tipo de episodios que se pueden definir a partir del sistema de variaciones padronales

Un episodio conceptualmente es un periodo durante el cual un individuo dado no ha modificado su permanecía a un estado concreto. Los episodios se definen, pues en base a un espacio de estados concreto y este puede ser conceptualizado de diferente manera en base a un objetivo concreto, por lo tanto es posible definir distinto tipo de episodios con un sistema de información concreto, véase por ejemplo la Ilustración 8 con alguno de los episodios que es posible definir dentro del subsistema de información padronal.

La consistencia temporal de una sucesión de suceso de altas y bajas es fácilmente visualizable si superponemos estos en un doble eje temporal, uno para las alta y otro para las bajas (Ilustración 9). Un sucesión de altas y bajas es temporalmente consistente si al dibujar los segmentos horizontales con comienzo en la diagonal a la fecha de alta y terminación en la vertical de la fecha de baja, el resultado es un gráfico perfectamente escalonado.

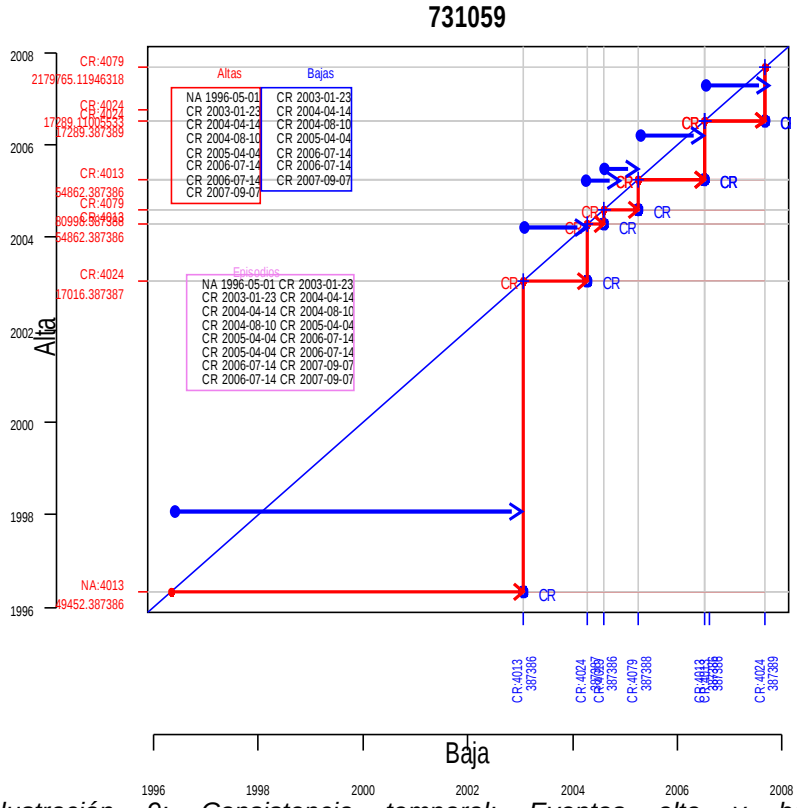


Ilustración 9: Consistencia temporal: Eventos alta y baja temporalmente ordenados en un doble eje

La visualización de la consistencia temporo-espacial requería incluir una tercera dimensión donde se representaría el espacio físico (Ilustración 10). La tercera dimensión representaría una determinada localización, un municipio, un portal o una vivienda. En cada caso el nivel de rigor exigido a la consistencia seria mayor. Una secuencia registral consiente se visualiza con un "**i**tinerario" formado por una sucesión continua de segmentos dirigidos (episodios y transiciones)

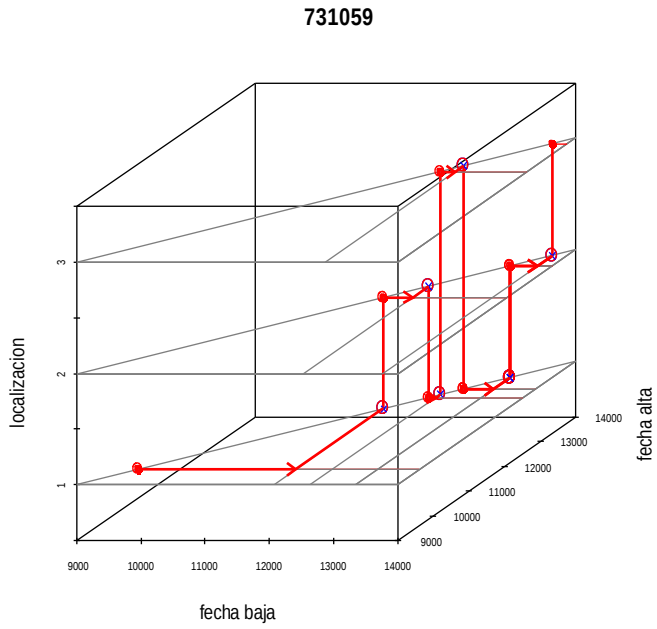


Ilustración 10: Consistencia espacio--temporal de una secuencia registral

Es evidente que no todas la secuencias registrarles son directamente coherentes, la transformación de una secuencia registral incoherente en otra coherente se realiza mediante técnicas de depuración e imputación de datos algunas de las cuales se describieron en una comunicación a las JECAS de Santander[2].

4.Esquema de entidades utilizado para gestionar la información del sistema

La información enumerada apartados previos se organiza en un esquema de base datos relacional, que brevemente se resume en la Ilustración 11.

En la actualidad, el sistema se estructura en tres sub-sistemas relativamente autónomos y que se corresponden mas o menos, con los circuitos de información preexistentes:

1.El sistema de eventos vitales (MNP) recogido a partir de los circuitos del Registro Civil

2.La información derivada de la operación censal de 2001.

3.El sistema de seguimiento continúo del estado residencial derivado del Padrón y que es la columna vertebral del sistema.

Los tres subsistemas se relacionan entre si por un conjunto de claves nominales que permiten la identificación (no segura) de las personas y las viviendas. Estas claves nominales (nombres, apellidos, direcciones postales…), son procesadas con objeto de obtener unos identificadores numéricos únicos de persona y vivienda (IDP, IDH).

Esta claves únicas (idp,idh) son asignadas internamente por el IEA mediante enlace de registros, que son periódicamente revisados, ya que somos conscientes de que el proceso de asignación no esta libre de errores. Su existencia permite la reconstrucción de las biografías individuales y de las relaciones de convivencia.

El núcleo del sistema es la información derivada del padrón de habitantes, ya que este nos suministra el estado vital y residencial de una persona concreta, en cualquier momento o periodo de seguimiento dado. Este subsistema nos dice si una persona esta viva y en que lugar físico (a nivel de vivienda) reside en un momento dado.

[2] Viciana F, Montañés V, Martínez D. Estadísticas de itinerarios demográficos a partir de la depuración de las variaciones padronales. Comunicación de las XVI JECAS en Santander.

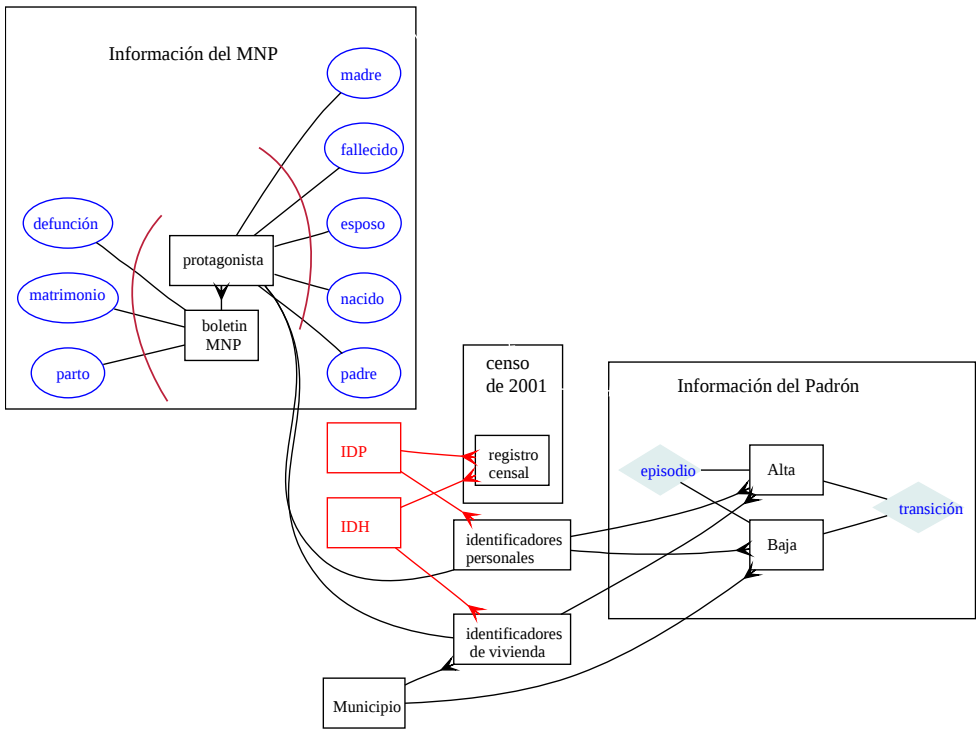
El subsistema "padrón" se estructura en dos tablas principales una de "altas" y otra de "bajas". La tabla de altas, nos informa básicamente del inicio de un nuevo estado (un nacimiento, una residencia en una nueva vivienda o la modificación de alguna característica personal) y la tabla de bajas implica la conclusión del estado anterior (por defunción, cambio de residencia o nueva modificación de alguna características personal) .

Las relaciones entre altas y bajas (en una biografía consistente, mismo "idp") se establecen en base a dos tipos de relaciones. Por un lado las relaciones sincrónicas (ocurren en el mismo tiempo, pero implican dos estados distintos que pueden ser dos residencias) y que determinan las "transiciones". Por otro lado las relaciones diacrónicas, ocurren en distinto tiempo, pero se aplican el mismo estado (alta y baja en la misma vivienda) estas relaciones definen los "episodios".

Para dar una idea del tamaño del subsistema padrón, entre 1996 y 2008 se han notificado mas de 25 millones de variaciones padronales (altas y bajas) que han afectado a mas de 9,5 millones de personas distintas, teniendo en cuenta que la población media de Andalucía en este periodo ha estado entorno a 8 millones de personas.

El subsistema MNP se estructura en base a dos tablas principales, la de los sucesos registrados en los boletines (parto, matrimonio y defunción) y las de los "protagonistas" de cada uno de estos sucesos, la persona física: nacido, madre, padre, fallecido y esposo. Son las personas físicas, a las que mediante sus identificadores personas, se le puede asignar un clave personal (idp). Igualmente las personas notifican una residencia a la cual se le intentara asignar su propia clave de vivienda (idh).

El subsistema censo de 2001 tiene gran interés, pues reporta gran cantidad de información sobre características personales no recogidas en los otros sistemas en el momento de la operación censal. Desafortunadamente, dadas las características del censo proyectado para España en el 2011 no va a ser posible renovar esta información en 2011. Se tiene previsto ir paulatinamente sustituyendo esta información por la derivada de registros administrativos.



En la actualidad los sistema de base de datos no están completamente adaptado al boceto de esquema objetivo expuesto previamente, pero se parece bastante. Las diferencias se dan en las tablas que recogen los identificadores individuales de personas y viviendas que de

momento se gestionan independientes en cada subsistema. El proceso de consolidación bastante avanzado para el caso de los identificadores individuales, y pensamos tenerlo concluido en este año. El tema de la consolidación y normalización de los de los identificadores de vivienda, estamos menos avanzados.

En la Ilustración 12 y 13 se muestra el estado actual del modelo de relacional de tablas que tenemos en funcionamiento para el subsistema Padrón y MNP y que ha permitido ya realizadas unas cuantas explotaciones, algunas de las cuales se muestran en estas mismas jornadas.

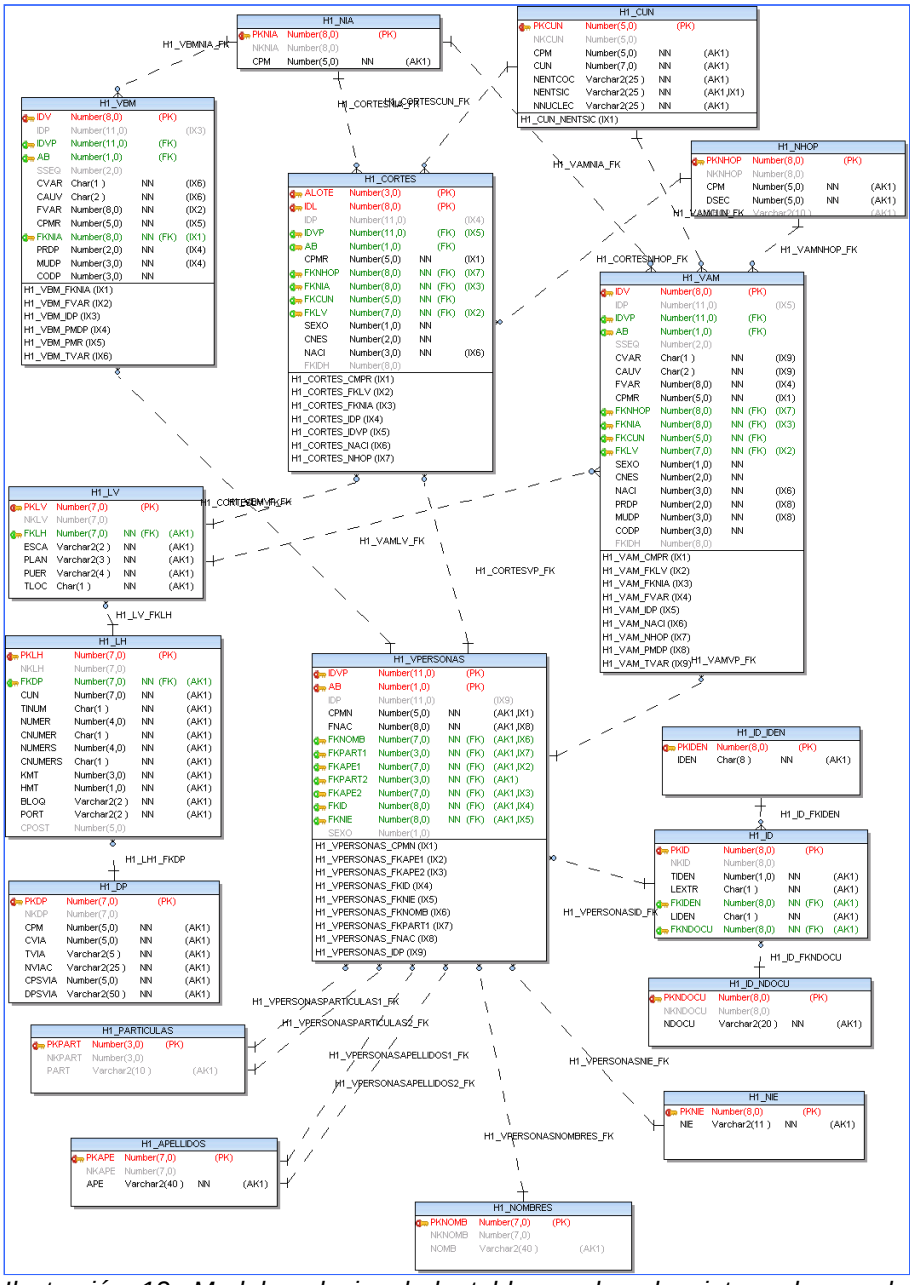


Ilustración 12: Modelo relacional de tablas padronales integradas en la BDLPA

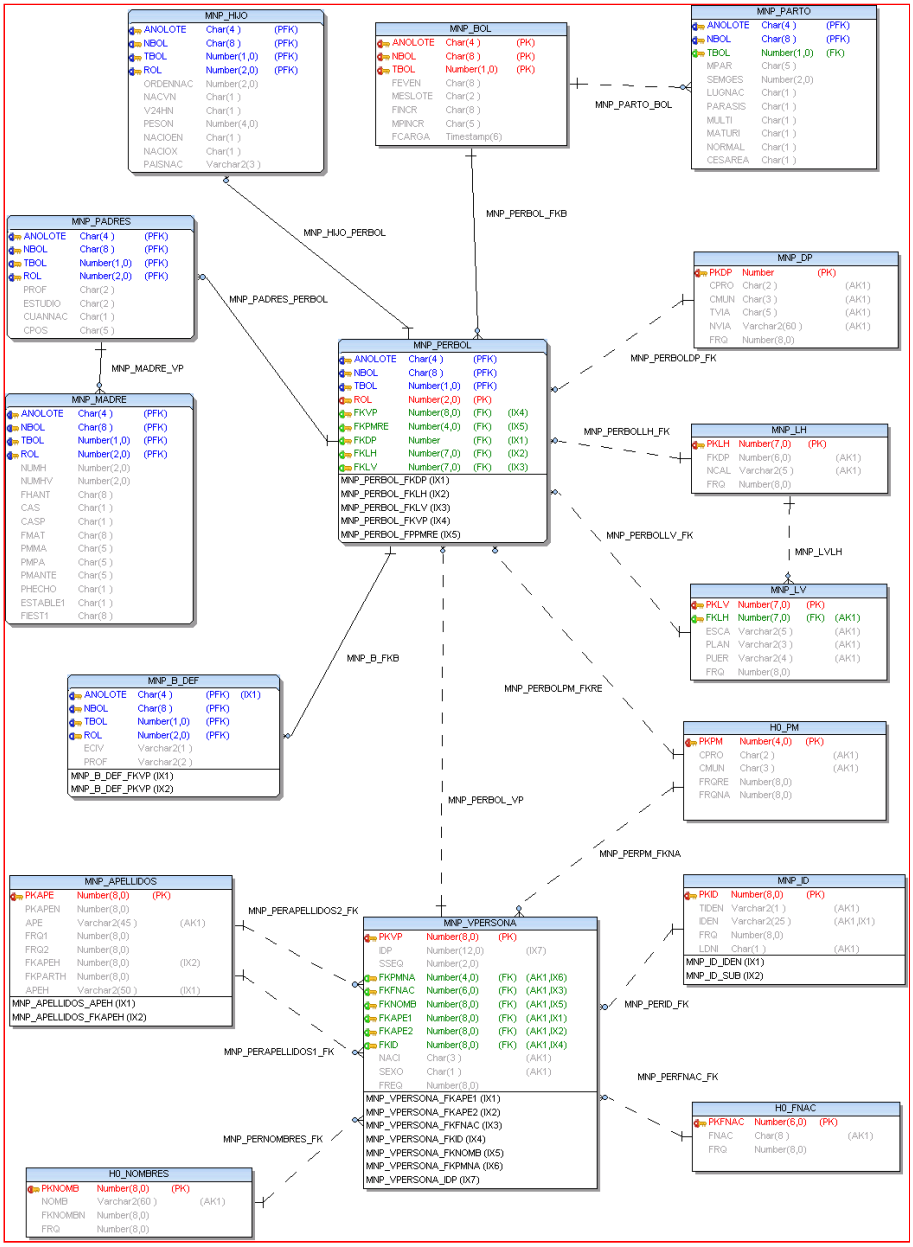


Ilustración 13: Modelo relacional de tablas del MNP integradas en la BDLPA