UBA – Facultad de Ciencias Exactas y Naturales – Departamento de Computación

Algoritmos y Estructura de Datos I

Segundo cuatrimestre de 2021

25 de Agosto de 2021

Trabajo Práctico de Especificación (TPE)

Análisis Habitacional Argentino

1. Modalidad de Trabajo

- El Trabajo Práctico se realiza en grupo de 3 personas. **NO SE ACEPTAN GRUPOS DE 1 PERSONA**
- Informar a la cátedra los integrantes del grupo para el 30/8.
- Fechas de Entrega:
  - **Primera Parte**: *Miércoles 8 de Septiembre - 23:55.*
  - **Segunda Parte**: *Viernes 1 de Octubre - 23:55.*
  - **Recuperatorio**: *Miércoles 1 de Diciembre - 23:55.*
- Entregable:
  1. Completar la solución de los ejercicios usando LATEX y siguiendo el template dado por la cátedra.
  2. Entregar la versión digital de la resolución en formato **.pdf** a través del campus de cada comisión por cualquiera de los miembros del grupo (realizar **una** sola entrega por grupo).

2. Introducción

La Ciencia de Datos es una actividad interdisciplinaria de gran auge en la actualidad, a veces conocida como aprendizaje estadístico (statistical learning), aprendizaje automático (machine learning) o minería de datos (data mining) entre otros. El análisis de datos es una tarea compleja que generalmente requiere la colaboración entre las Ciencias de la Computación, la Matemática y las ciencias empíricas especializadas en los diversos objetos de estudio (Ciencias Sociales, Biología, Química, Física).

El Instituto Nacional de Estadística y Censos (INDEC[1]) realiza a nivel nacional La Encuesta Permanente de Hogares (EPH). Las encuestas registran alrededor de 80 atributos de los hogares y 150 atributos de las personas que los habitan. Los hogares corresponden a lugares estratégicamente asignados para obtener una muestra lo más representativa posible de la población que habita el territorio argentino. Los datos de esta encuesta están divididos en dos tablas: la tabla **hogares** y la tabla **personas**.

En los Trabajos Prácticos de Especificación y de Implementación vamos a realizar un análisis de la base de datos públicos de la Encuesta Permanente de Hogares recolectado por el INDEC. En este trabajo práctico vamos a utilizar el lenguaje de especificación de la materia, basado en la lógica de primer orden, para definir con precisión los problemas que vamos a resolver en el Trabajo Práctico de Implementación. ¡Veamos la base de datos!

3. Base de datos

La Encuesta Permanente de Hogares se realiza de forma trimestral y se comparte en formato estandarizado de tipo texto o Excel[2]. Las filas corresponden a las unidades de análisis (hogares o personas según la tabla) y las columnas corresponden a los atributos de estas unidades de análisis. Los datos de las tablas, para este trabajo, corresponden al tipo entero. Todos los atributos categóricos se codifican con tipo de datos enumerados. Los hogares se identifican con un código numérico llamado

[1] `https://www.indec.gob.ar/`

[2] `https://www.indec.gob.ar/bases-de-datos.asp`

CODUSU, que es único para cada hogar. Los individuos se identifican con un par CODUSU-COMPONENTE, siendo el primer código el correspondiente al CODUSU del hogar donde vive el individuo y el segundo un número que identifica los individuos dentro de un mismo hogar. Veamos con qué atributos contamos.

3.1. Tabla HOGARES

- HOGCODUSU: Identificador o clave (IN) del hogar. Además permite hacer el seguimiento a través de los trimestres.
- HOGAÑO: Año de relevamiento.
- HOGTRIMESTRE: Trimestre del año de relevamiento.
- HOGLATITUD: Latitud (geolocalización del hogar).
- HOGLONGITUD: Longitud (geolocalización del hogar).
- II7: Régimen de tenencia de los habitantes:
  - 1 - Propietario
  - 2 - Inquilino
  - 3 - Ocupante
- REGION: Código de Región:
  - 1 - Gran Buenos Aires
  - 2 - NOA
  - 3 - NEA
  - 4 - Cuyo
  - 5 - Pampeana
  - 6 - Patagonia
- MAS_500: El hogar se ubica dentro de un aglomerados de más de 500.000 habitantes:
  - 0 - NO
  - 1 - SI
- IV1: Tipo de hogar (por observación):
  - 1 - Casa
  - 2 - Departamento
  - 3 - Pieza de inquilinato
  - 4 - Pieza en hotel/pensión
  - 5 - Local no construido para habitación
- IV2: Cantidad total de ambientes o habitaciones (sin contar baño/s, cocina, pasillo/s, lavadero, garage).
- II2: De esos, ¿cuántos usan habitualmente para dormir?
- II3: ¿Utiliza alguno exclusivamente como lugar de trabajo (para consultorio, estudio, taller, negocio, etc.)?
  - 1 - Si
  - 2 - No

3.2. Tabla PERSONAS

- INDCODUSU: Identificador único o clave (IN) del hogar. Se corresponde a un HOGCODUSU de la tabla hogares. Además permite hacer el seguimiento a traés de los trimestres.
- COMPONENTE: Número de orden que identifica a una persona dentro de un hogar.
- INDAÑO: Año de relevamiento.
- INDTRIMESTRE: Trimestre del año de relevamiento.
- CH4: Género:
  - 1 - Varón

- 2 - Mujer

- `CH6`: Cuantos años cumplidos tiene.
- `NIVEL_ED`: Estudios universitarios completos:
  - 0 - NO
  - 1 - SI
- `ESTADO`: Condición de actividad:
  - 0 - Desocupado, Inactivo
  - 1 - Ocupado
  - -1 - No informado
- `CAT_OCUP`: Categoría ocupacional (Para ocupados y desocupados con ocupación anterior):
  - 0 - Ns./Nr.
  - 1 - Patrón
  - 2 - Cuenta propia
  - 3 - Obrero o empleado
  - 4 - Trabajador familiar sin remuneración
- `p47T`: Monto de ingreso total individual. Puede ser -1 si no fue informado.
- `PP04G`: Dónde realiza principalmente sus tareas:
  - 1 - En un local / oficina / establecimiento negocio / taller / chacra / finca
  - 2 - En puesto o kiosco fijo callejero
  - 3 - En vehículos: bicicleta / moto / autos / barcos / botes (no incluye servicio de transporte)
  - 4 - En vehículo para transporte de personas y mercaderías-aéreos, marítimo, terrestre (incluye taxis, colectivos, camiones, furgones, transporte de combustible, mudanzas, etc.)
  - 5 - En obras en construcción, de infraestructura, minería o similares
  - 6 - En este hogar
  - 7 - En el hogar del socio o del patrón
  - 8 - En el domicilio / local de los clientes
  - 9 - En la calle / espacios públicos / ambulante / de casa en casa / puesto móvil callejero
  - 10 - En otro lugar

3.3. Tipos:

type $dato = \mathbb{Z}$

type $individuo = seq\langle dato \rangle$

type $hogar = seq\langle dato \rangle$

type $eph_i = seq\langle individuo \rangle$

type $eph_h = seq\langle hogar \rangle$

type $joinHI = seq\langle hogar \times individuo \rangle$

Tanto *individuo* como *hogar* se pueden pensar como una línea en las matrices eph_i y eph_h.

3.4. Acceso a las columnas de las tablas

Con el objetivo de crear una manera dinámica para el acceso a las columnas, y no preocuparnos desde la especificación de la ubicación de los ítems en la tabla, utilizaremos el tipo enumerado para los ítems utilizando el código correspondiente: Definimos los tipos de *items* así:

```
enum ItemIndividuo {
    INDCODUSU, INDAÑO, INDTRIMESTRE, COMPONENTE,  NIVEL_ED, ESTADO, ...
}

enum ItemHogar {
    HOGCODUSU, HOGAÑO, HOGTRIMESTRE, ...
}
```

- Valen:
  - ord(ESTADO) = 5
  - ItemIndividuo(0) = INDCODUSU
  - ItemHogar(0) = HOGCODUSU
- Además:
  - ord(ALGO) = $\perp$, si ALGO no está en Item
  - Item(n) = $\perp$, si n es negativo o es mayor o igual que la cantidad de enumerados de Item.
- Sugerencia, definir
  - `aux @estado` : $\mathbb{Z} = ord(ESTADO)$`;`
    En este ejemplo, para acceder a la columna ESTADO de un individuo, se puede escribir

    $(\forall p : individuo)\ p[@estado]$ ...

4. EJERCICIOS

4.1. Primera Parte

Especificar los siguientes problemas:

1. **proc esEncuestaVálida**(**in** $th : eph_h$, **in** $ti : eph_i$, **out** $result$: Bool).
   El procedimiento devuelve verdadero si se verifica:
   - Que th y ti son matrices, es decir, que en el interior de cada una, todos sus vectores tienen la misma longitud
   - Que existe al menos un hogar en th y un individuo en ti
   - Que la cantidad de columnas (tamaño de los vectores) es igual a la cantidad variables de la tabla (o de enumerados de **Item**)
   - Que los hogares tienen individuos asociados y viceversa, es decir, que no hay individuos sin hogares ni hogares sin individuos
   - Que no hay individuos ni hogares repetidos
   - Que el año y trimestre de relevamiento es el mismo para todos los registros
   - Que la cantidad de miembros del hogar es menor o igual a 20
   - Que el atributo `IV2` es mayor o igual al atributo `II2`
   - Que todos los atributos categóricos tienen valores en el rango esperado. Por ejemplo, el atributo `REGION` sólo debería tener valores entre 1 y 6 inclusive
2. **proc histHabitacional**(**in** $th : eph_h$, **in** $ti : eph_i$, **in** $region : \mathbb{Z}$, **out** $res : seq\langle\mathbb{Z}\rangle$) .
   Dada una encuesta válida, se desea construir el histograma habitacional correspondiente a una región dada como parámetro de entrada. El histograma se representa con una secuencia de enteros **res** donde la i-ésima posición contiene la cantidad de hogares de tipo casa con i habitaciones en la `región` recibida por parámetro. El largo de la secuencia de salida depende de la máxima cantidad de habitaciones en la región.
3. **proc laCasaEstaQuedandoChica**(**in** $th : eph_h$, **in** $ti : eph_i$, **out** $res : seq\langle\mathbb{R}\rangle$) .
   Dada una encuesta válida, se pide calcular por cada una de las 6 regiones argentinas, la proporción de hogares tipo casa con hacinamiento crítico. Los hogares con hacinamiento crítico son aquellos en los cuales hay en promedio más de tres personas por cuarto. Las casas deben estar ubicadas en aglomeraciones de menos de 500.000 habitantes. Agrupar los cálculos por región en una secuencia ordenada de acuerdo al código de la columna `REGION`.
4. **proc creceElTeleworkingEnCiudadesGrandes**(**in** $t1h : eph_h$, **in** $t1i : eph_i$, **in** $t2h : eph_h$, **in** $t2i : eph_i$, **out** res : Bool)
   Dadas dos encuestas válidas, detectar si hay un incremento (en proporción) interanual del Teleworking en ciudades de más de 500.000 habitantes. Para ello calcular la proporción de individuos que realizan sus tareas laborales en su hogar (`PP04G`), entre dos encuestas de años diferentes, pero del mismo trimestre. Específicamente, t1h y t1i son anteriores a t2h y t2i. Verificar solo para hogares tipo casas o departamentos. Para simplificar el análisis, se considerarán como haciendo Teleworking, los hogares que tengan ambientes reservados para el trabajo.
5. **proc costoSubsidioMejora**(**in** $th : eph_h$, **in** $ti : eph_i$, **in** $monto : \mathbb{Z}$, **out** $res : \mathbb{Z}$) .
   Dada una encuesta válida, y un monto de subsidio se desea calcular el costo total de implementar un subsidio de mejora habitacional para aquellos hogares que sean casas de tenencia propia y cuya cantidad de habitaciones que utilizan para dormir (atributo II2) sea estrictamente inferior a la cantidad de habitantes menos dos.

4.2. Segunda Parte

Especificar los siguientes problemas:

6. **proc generarJoin**(**in** $th : eph_h$, **in** $ti : eph_i$, **out** $junta : joinHI$) .
   Dada una encuesta válida, devuelve la combinación de las tables de hogares e individuos generando la tabla "junta"tal que cada fila es una 2-upla de la cual el primer elemento es de tipo hogar y el segundo de tipo individuo que satisfacen la condición que el valor en HOGCODUSU del primero es igual al valor de INDCODUSU del segundo.
7. **proc ordenarRegionYTipo**(**inout** $th : eph_h$, **inout** $ti : eph_i$) .
   Dada una encuesta válida, ordenar la encuesta de hogares `th` de acuerdo a: 1) El código de región: todos los del gran buenos aires primero, luego los de NOA y así sucesivamente (siguiendo el orden dado por el número de categoría). 2) Dentro de cada región, ordenar de forma creciente por `CODUSU`.

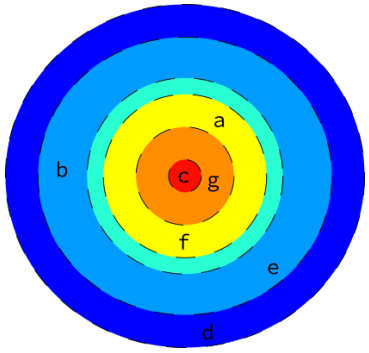
   Además ordenar la encuesta individuos `ti`, según: 1) El `CODUSU` de th luego de realizar el ordenamiento, 2) Dentro del mismo hogar, ordenar por `COMPONENTE` de menor a mayor.

8. **proc muestraHomogenea**(**in** $th : eph_h$, **in** $ti : eph_i$, **out** $res : seq\langle hogar \rangle$) .
Dada una encuesta válida, encontrar una secuencia de hogares lo más larga posible tal que la diferencia de ingresos totales sea la misma para cada par de hogares consecutivos. Debe estar ordenada de menor a mayor por cantidad de ingresos. De no encontrarse una secuencia de al menos 3 elementos, devolver una secuencia vacía.

9. **proc corregirRegion**(**inout** $th : eph_h$, **in** $ti : eph_i$) .
Dada una encuesta válida, se desea agrupar las regiones de Gran Buenos Aires y Pampeana. Para cada hogar de Gran Buenos Aires, cambiar la región del hogar a Pampeana. Todo lo demás deberá permanecer igual.

10. **proc histogramaDeAnillosConcentricos**(**in** $th : eph_h$, **in** $centro : \mathbb{Z} \times \mathbb{Z}$, **in** $distancias : seq\langle \mathbb{Z} \rangle$, **out** $result : seq\langle \mathbb{Z} \rangle$) .
Dada una tabla de hogares, un punto central ("centro") definido como un par (latitud, longitud) y una lista no vacía y estrictamente creciente de distancias no nulas al centro, devuelve otra lista que contiene la cantidad de hogares que se encuentran en los anillos concéntricos determinados por las distancias[3] con respecto al punto central.

    Para ilustrar la situación, consideremos el siguiente gráfico:



    - El "centro" es el punto central de las circunferencias
    - Las "distancias" son los radios de las circunferencias
    - Y "result" contiene:
        - en la primer posición: la cantidad de hogares que están contenidas en la circunferencia más chica (roja)
        - en las posiciones posteriores: la cantidad de hogares contenidas en cada anillo de color.

    Para este ejemplo, donde las letras indican esquemáticamente las ubicaciones geográficas de los hogares presentes, y donde un posible valor de *distancias* puede ser $[1, 3, 5, 6, 9, 11]$, el valor esperado de *result* es $[1, 1, 2, 0, 2, 1]$.

11. **proc quitarIndividuos**(**inout** $th : eph_h$, **inout** $ti : eph_i$, **in** $busqueda : seq\langle (ItemIndividuo, dato) \rangle$, **out** $result : (eph_h, eph_i)$)
Dada una encuesta válida y una búsqueda de individuos válida, se desea quitar los individuos que coinciden con todos los términos de búsqueda, y en base a este resultado se desea también quitar sus hogares correspondientes (comparando el `CODUSU`).

    Una búsqueda se define como una lista de pares ordenados $(item : ItemIndividuo, valor : dato)$, y para que sea válida debe ocurrir que el *item* sea un valor válido de *ItemIndividuo*, sin repetirse en la búsqueda; y *valor* es el valor que se desea que el individuo tenga para su *item* asociado.

    Los individuos y hogares quitados deben ser devueltos en *result*.

    Ejemplo:

    quitarIndividuos*(TH, TI, ⟨(INDAÑO, 2020),(REGION,5)⟩, result)* debería devolver en *result* tablas de hogares e individuos que contengan solamente registros de 2020 de la región pampeana; mientras que TH y TI deberán contener el resto de los registros originales.

[3] para calcular la distancia entre un hogar y el "centro" hay que considerar la distancia *euclidiana*, es decir, la distancia entre dos coordenadas c_1 y c_2 es igual a la raíz cuadrada de la suma de los cuadrados de sus diferencias en latitud y longitud.