Identificació i protecció de dades tabulares: el cas de l’Estadística sobre pensions de l’INSS

Elisenda Vila Jofre

Director: Enric Ripoll
Tutor: Manel Falguera
Dep. Econometria, Estadística i Economia Espanyola
El qui coneix els homes és intel·ligent.
El qui es coneix a sí mateix és savi.
El qui venç els altres és vigorós.
El qui es venç a sí mateix és veritablement fort.
El qui s’accontenta amb el que té és ric.
El qui persevera és home de voluntat.
El qui no perd el seu Centre resisteix.
El qui mor i no pereix té llarga vida.

Lao Tse
Resum

Paraules clau: estadística oficial, control de la revelació estadística, pensions contributives, problema de supressió de cel·les

Un dels principals serveis que presten els Instituts d’estadística és la disseminació de gran quantitat de dades tabulars, que s’obtenen a partir del creuament de una o diverses variables categòriques d’un fitxer de microdades.

Els Instituts d’estadística han de garantir que no es pot obtenir informació confidencial individual a partir de les dades tabulars publicades. El control de la revelació estadística apareix, doncs, com un camp en constant evolució que permet dotar als Instituts d’estadística de les eines necessàries per a assolir l’equilibri entre la necessitat d’informació i la necessitat de protecció de la privació dels informants.

Aquest projecte es centra en l’aplicació de les tècniques de control de la revelació estadística en el cas català a partir de les dades de pensions contributives de la Seguretat Social.

La primera part del projecte descriu els coneixements previs necessaris per a entendre plenament el cas. El primer capítol descriu les particularitats de l’estadística oficial, així com l’estructura del sistema estadístic vigent a Catalunya. El segon capítol es centra en les normatives aplicables en el cas de la protecció de dades personals i en la figura del secret estadístic. El tercer capítol està dedicat a les nomenclatures i classificacions estadístiques, una peça clau en la definició de les categories de les variables en l’estadística oficial. El capítol quart introduceix els conceptes relatius a la temàtica de les dades a protegir: les pensions contributives de la Seguretat Social. Al capítol cinquè es revisa la difusió actual de les dades de pensions a l’Estat espanyol i es defineixen els objectius de difusió marcats. Al capítol sisè s’introdueixen els conceptes relatius al control de la revelació estadística en el seus tres grans camps d’aplicació: les microdades, les dades tabulars i l’anàlisi de resultats.

La segona part del projecte tracta ja de manera concreta el cas de la protecció de les dades tabulars procedents del fitxer de pensions contributives de la Seguretat Social. El capítol setè descriu les decisions prèvies a la protecció de les taules. Les dades tabulars s’han protegit resolent el problema de la supressió de cel·les o CSP. Al capítol vuitè es descriuen amb detall les tècniques usades per a resoldre el CSP. El capítol nou està dedicat a descriure les característiques del software usat: el paquet estadístic sdcTable per a R. Al capítol desè es presenten els resultats obtinguts. El capítol onze es presenten les conclusions del projecte: tant pel que respecte a la difusió segura dels resultats, com a la possible inclusió d’aquestes tècniques i aquest software en altres softwares estadístics d’ús habitual.
Abstract

Keywords: official statistics, statistical disclosure control, pensions, cell supression problem

One of the main services provided by National Statistical Institutes (NSIs) is the dissemination of large amount of tabular data, obtained by crossing one or several categorical variables belonging to a microdata file.

National Statistical Institutes must guarantee that no individual confidential information can be derived from the published tabular data. Statistical Disclosure Control appears as a field in constant evolution that furnishes NSIs with the necessary tools to achieve the balance between the need of information and the need to protect respondents privacy.

This project focuses on the application of Statistical Disclosure Techniques in the Catalan case with the use of data about the Social Security pensions.

The first part of the project describes the previous knowledge necessary to fully understand the case. The first chapter deals with the peculiarities of official statistics, as well as the Catalan Statistical System structure. The second chapter focuses on the regulations concerning personal data protection and statistical secret. The third chapter is dedicated to nomenclatures and statistical classifications, a key point in defining variable categories in official statistics. The fourth chapter introduces concepts related to the specific subject of the data: Social Services pensions. In chapter number five the actual dissemination of pension data in Spain is revised and the dissemination objectives are set. In chapter number six, concepts regarding Statistical Disclosure Control techniques in the three main application areas are introduced: microdata, tabular data and output checking.

The second part of the project deals in a more detailed way with the protection of tabular data from Social Society pensions file. Chapter number six describes the decisions taken prior to the tabular data protection. Tables have been protected by solving the Cell Suppression Problem (CSP). In chapter eight there is an in depth description of the techniques used to solve the CSP problem. Chapter nine outlines the used software characteristics: the statistical package sdcTable for R. Chapter ten contains the results. In chapter eleven we can find the project conclusions: those related to the safe dissemination of data, as well as those related to the possible inclusion of these techniques and this software in other frequently used statistical software.
Índex general

Índex de figures iii

Índex de taules v

Part 1. Conceptes previs 1

Capítol 1. L’Estadística Oficial 3
   1. Principis fonamentals de l’estadística oficial 3
   2. Organització del sistema estadístic a Catalunya 4
   3. El sistema estadístic europeu 6
   4. El sistema estadístic de l’Administració General de l’Estat 9
   5. El sistema estadístic català 11

Capítol 2. Normativa aplicable a la protecció de dades 15
   1. Normativa relativa a la protecció de dades de caràcter personal 15
   2. Normativa relativa al secret estadístic 17

Capítol 3. Nomenclatures i classificacions estadístiques 21
   1. Característiques i conceptes clau 21
   2. Obligatorietat d’ús de les classificacions 26

Capítol 4. Estadística de pensions de la Seguretat Social 29
   1. Tipus de pensions contributives 29
   2. Règims d’afiliació a la Seguretat Social 36
   3. Recomanacions europees sobre estadística de pensions 37

Capítol 5. La difusió de resultats de l’Estadística de Pensions 41
   1. Ministeri de Treball i Seguretat Social 41
   2. Instituts d'Estadística de la resta de Comunitats Autònones 42
   3. Objectius de difusió 45

Capítol 6. El Control de la Revelació Estadística 49
   1. Mètodes de control de la revelació estadística per a microdades 50
   2. Mètodes de control de la revelació estadística per a dades tabulars 58
   3. Mètodes de control de la revelació estadística per l’anàlisi de resultats 66

Part 2. Resolució del cas 69

Capítol 7. Decisions preliminars 71
ÍNDICE GENERAL

1. Elección de les dades sobre les que aplicar els mètodes 71
2. Elecció dels mètodes de protecció 72
3. Elecció del software 72
4. Pre-proces de les microdades 73

Capítol 8. El problema de supressió de cel·les (CSP) 75
1. Definició del problema 75
2. Descomposició de Benders per al CSP 77
3. El mètode de l’hipercub 80
4. El mètode HiTas 82

Capítol 9. El paquet sdcTable 87
1. Característiques del paquet sdcTable 87
2. Les regles disponibles per a realitzar supressions primàries 89
3. Estats d’anonimització de les cel·les 89
4. Els mètodes disponibles de resolució 89

Capítol 10. Resultats de l’aplicació dels mètodes de resolució del CSP 95
1. Característiques de l’ordinador on s’han realitzat els càlculs 95
2. Creació del problema amb la funció makeProblem 95
3. Resultats per a dades a nivell d’àmbit territorial 97
4. Resultats per a dades a nivell municipal 98

Capítol 11. Conclusions 103
1. Objectius de difusió segura 103
2. Integració al sistema de producció i difusió 103
3. Millora del temps d’execució 105
4. Incidències amb el software 105
5. Futures accions 106

Apèndix A. El mètode SIMPLEHEURISTIC 107
1. El codi original 107
2. El codi modificat 111

Apèndix B. La funció protectTable 113

Apèndix C. Exemple de resum de la solució del CSP 115

Apèndix. Índex alfabètic 117

Apèndix. Bibliografía 119
Índex de figures

1.1 Aproximació a l’Estadística Oficial 3
1.2 Interdependències entre els sistemes estadístics 5

4.1 Prestacions de la Seguretat Social 30
4.2 Règims de la Seguretat Social 37

5.1 Estructura del pla de tabulació 46

6.1 Informació vs. privadesa 50
6.2 Gràfic Risc – Utilitat 51

8.1 Estructura jeràrquica de la variable TP 83
8.2 Estructura jeràrquica de la variable R 84

9.1 Funcions visibles als usuaris en el paquet sdcTable 88

10.1 Resultats del CSP per dades a nivell d’àmbit 99
10.2 Municipis no publicables de la taula Municipi per tipus de pensió i sexe 101
10.3 Municipis no publicables de la taula Municipi per règim i sexe 102
10.4 Municipis no publicables de la taula Municipi per règim, tipus de pensió i sexe 102
Índex de taules

5.1 Difusió de l’Estadística de Pensions a la resta de Comunitats Autònomes 45

6.1 Mètodes no pertorbatius segons el tipus de dades 53
6.2 Mètodes pertorbatius segons el tipus de variables. ’x’ significa aplicable i ’(x)’ significa només aplicable a variables categòriques ordinals 54
6.3 Ocupació segons edat al municipi. Exemple: el cas del metge del poble 58
6.4 Mitjana del sou segons ocupació i edat al municipi. Exemple: el cas del metge del poble 59
6.5 Principals resultats. Exemple: el cas del gran grup empresarial 59
6.6 Principals resultats. Empresa E 60
6.7 Tècniques de protecció de a dades tabulars 62
6.8 Supressió de cèlles 63
6.9 Protecció per intervals 64
6.10 Arrodoniment controlat 65
6.11 Ajust controlat 65
6.12 Mètodes de control de la revelació estadística per anàlisi de resultats. Tipus de resultats. 67

8.1 Les classes definides al creuar R amb TP 85

10.1 Temps de creació de l’objecte sdcProblem usant microdades. 96
10.2 Temps de creació de l’objecte sdcProblem usant dades ja tabulades. El temps representat inclou el temps de càlcul de les taules en el format apropriat. 96
10.3 Resultats del CSP per dades a nivell d’àmbit 98
10.4 Resultats del CSP per dades a nivell de municipi 100
Part 1

Conceptes previs
Capítol 1
L’Estadística Oficial

Si cerquem una definició única de què és l’estadística oficial ens trobarem amb un greu problema, ja que cada estat descriu de manera lleugerament diferent aquest mateix concepte. Malgrat tot, d’una manera molt genèrica, podríem definir l’estadística oficial com aquella branca de l’estadística, normalment exercida pels estats, que disposa d’un marc conceptual estandarditzat i està regulada per un marc jurídic propi. És precisament l’existència d’un marc jurídic propi el que impossibilita una definició única del concepte.

Figura 1.1. Aproximació a l’Estadística Oficial

Malgrat no disposem de lleis mundials al respecte, sí que a través de la divisió estadística de les Nacions Unides (la més alta autoritat estadística del món) s’adopten al 1994 els Principis fonamentals de l’estadística oficial.

1. Principis fonamentals de l’estadística oficial

Els principis fonamentals de l’estadística oficial es sintetitzen en 10 punts.

(1) Les estadístiques oficials constitueixen un element indispensable en el sistema d’informació d’una societat democràtica i proporcionen al Govern, a l’economia i al públic dades sobre la situació econòmica, demogràfica, social i ambiental. Amb aquesta finalitat, els organismes oficials d’estadística han de
1. L’ESTADÍSTICA OFICIAL

compilar i facilitar de forma imparcial estadístiques oficiales de comprovada utilitat pràctica per què els ciutadans puguin exercir el seu dret a mantenir-se informats;

(2) Per mantenir la confiança en les estadístiques oficials, els organismes estadístics han de decidir, a partir de consideracions estrictament professionals, inclosos els principis científics i la ètica professional, sobre els mètodes i procediments per a la reunió, processament, emmagatzematge, i la presentació de dades estadístiques;

(3) Per a facilitar una interpretació correcta de les dades, els organismes estadístics han de presentar informació conforme a normes científiques sobre les fonts, mètodes i procediments estadístics;

(4) Els organismes estadístics tenen dret a formular observacions sobre interpretacions errònies i la utilització indeguda de les estadístiques;

(5) Les dades per a finalitats estadístiques es poden obtenir de tot tipus de fonts, ja sigui enquestes estadístiques o registres administratius. Els organismes estadístics han de seleccionar la font segons la qualitat, la oportunitat, el cost i la càrrega que suposaran;

(6) Les dades que reuneixin els organismes estadístics per a la compilació estadística, tan si es refereixen a persones físiques com a persones jurídiques, han de ser estrictament confidencials i s’han d’usar exclusivament amb finalitats estadístiques;

(7) S’han de donar a conèixer al públic les lleis, reglaments i mesures que regeixen l’operació dels sistemes estadístics;

(8) La coordinació entre els organismes estadístics a nivell nacional és indispen-sable per a aconseguir la coherència i l’eficiència del sistema estadístic;

(9) La utilització per part dels organismes estadístics de cada país de conceptes, classificacions i mètodes internacionals fomenta la coherència i eficiència dels sistemes estadístics a nivell oficial;

(10) La cooperació bilateral i multilateral en l’esfera de l’estadística contribueix a millorar els sistemes estadístics oficials en tots els països.

Els principis fonamentals de l’estadística oficial descriuen els dos trets essencials que defineixen aquesta àrea de l’estadística: l’existència d’un marc conceptual estandarditzat i la regulació a partir de lleis i reglaments.

2. Organització del sistema estadístic a Catalunya

A Catalunya hi ha la particularitat que es juxtaosplen tres sistemes estadístics. D’una banda el Sistema Estadístic Europeu, de l’altra el Sistema Estadístic Espa

nyol que forma part del primer i finalment el Sistema Estadístic de Catalunya, que és un sistema propi però que a la vegada participa de l’estatal.

En la Figura 1.2 es poden veure representades les interdependències entre els tres sistemes estadístics.
2. ORGANITZACIÓ DEL SISTEMA ESTADÍSTIC A CATALUNYA

Figura 1.2. Interdependències entre els sistemes estadístics
3. El sistema estadístic europeu

Des que l’any 1958 es crea la Comunitat Econòmica Europea ja es visualitza la necessitat de basar els plantejaments i implementacions de polítiques Comunitàries en dades estadístiques sòlides i comparables. Així es crea, de manera gradual, el Sistema Estadístic Europeu (ESS) amb l’objectiu de proporcionar estadístiques comparables a nivell de la Unió Europea.

Eurostat és l’autoritat estadística Comunitària. Es crea l’any 1953 per a assolir els requeriments de la Comunitat Europea del Carbó i de l’Acer (CECA), tot i que no adopta aquest nom fins l’any 1959.

El Sistema Estadístic Europeu està compost per Eurostat, pels instituts d’estadística nacionals dels Estats membres de la Unió, així com dels estats membres de l’Associació Europea de Lliure Comerç (EFTA) i els països candidats a entrar a la Unió. Com a òrgan consultiu està present el Comitè del Sistema Estadístic Europeu. També formen part de l’ESS el Banc Central Europeu i la Divisió Estadística de la Comissió Econòmica per Europa de les Nacions Unides.

El Reglament 223/2009 regula l’estadística europea i deroga els reglaments i decisions anteriors al respecte. La Decisió 2012/504/UE sobre Eurostat estén aquest reglament, tot i que no aporta novetats importants.

Característiques principals:

(1) El programa estadístic determina quines són les estadístiques europees, és a dir, aquelles estadístiques que es consideren necessàries per desenvolupar l’activitat de la Comunitat. S’adopta el Codi de Bones Pràctiques de les estadístiques europees.

(2) L’elaboració i difusió de les estadístiques europees es regueix per 6 principis estadístics, que es desenvolupen en el codi de Bones Pràctiques:
   (a) Independentència professional
   (b) Imparcialitat
   (c) Objectivitat
   (d) Fiabilitat
   (e) Secret estadístic
   (f) Rendibilitat

(3) El Comitè del Sistema Estadístic Europeu és l’òrgan consultiu encarregat del desenvolupament, seguiment i establiment de prioritats en el programa estadístic europeu. També és l’organisme de referència per a les consultes de secret estadístic i per l’elaboració del Codi de Bones Pràctiques en les estadístiques europees, i en general per a qualsevol dubte de cair eminentment metodològic al crear o aplicar els programes estadístics.

(4) Per a garantir la qualitat dels resultats de les estadístiques produïdes s’apliquen els següents criteris de qualitat:
   (a) Pertinència
   (b) Precisió
   (c) Actualitat
   (d) Puntualitat
   (e) Accessibilitat i claredat
3. EL SISTEMA ESTADÍSTIC EUROPEU

(f) Comparabilitat
(g) Coherència

(5) El programa estadístic europeu proporciona el marc per desenvolupar, elaborar i difondre les estadístiques europees, els principals àmbits i els objectius de les mesures previstes durant un període no superior a cinc anys. S’estableix per el Parlament Europeu i per Eurostat. Es presenten un informe intermedi i un informe final d’avaluació que s’eleven al Parlament Europeu i al Consell. El programa estadístic europeu es concreta en els programes de treball anual.

(6) Les dades sobre unitats estadístiques individuals poden difondre’s en forma de fitxer d’ús públic consistent en registres anònims elaborats de manera que no sigui identifiable cap unitat estadística, ni de forma directa ni indirecta, prenent en consideració tots els mitjans pertinents que un tercer pugui usar raonablement.

(7) Les dades confidencials obtingudes exclusivament per a la producció d’estadístiques europees podran ser usades pels instituts nacionals d’estadística i altres autoritats nacionals i per Eurostat únicament amb finalitats estadístiques.

(8) Els instituts nacionals d’estadística i altres autoritats nacionals i Eurostat adoptaran totes les mesures reglamentàries, administratives, tècniques i d’organització necessàries per garantir el control de la revelació estadística. S’adoptaran també totes les mesures necessàries per a garantir l’harmonització de principis i directrius referents a la protecció física i lògica de les dades confidencials.

(9) Tot el personal dels instituts nacionals d’estadística i altres autoritats nacionals que tinguin accés a dades confidencials estan obligats a respectar aquesta confidencialitat fins i tot després d’haver cessat en les seves funcions.

(10) Eurostat o els instituts nacionals d’estadística o altres autoritats nacionals poden concedir accés a dades confidencials que només permetin la identificació indirecta de les unitats estadístiques a investigadors que portin a terme anàlisis estadístics amb finalitats científiques, segons les modalitats, normes i condicions d’accés comunitàries.

(11) Els Estats membres i la Comissió prendran mesures apropriades per prevenir i sancionar qualsevol violació del secret estadístic.

3.1. Codi de Bones Pràctiques de les Estadístiques Europees

La primera versió del Codi de Bones Pràctiques de les Estadístiques Europees és adoptat per Comitè del Programa Estadístic el 24 de febrer de 2005 i promulgat en la recomanació de la Comissió del 25 de maig de 2005 relativa a la independència, la integritat i la responsabilitat de les autoritats estadístiques dels Estats membres i de la comunitat. Aquesta primera versió s’estructura en 15 principis desenvolupats en diversos punts.

A partir del Reglament 223/2009 el Codi de Bones Pràctiques de les Estadístiques Europees ja no es menciona com a una recomanació sinó que s’afirma que les estadístiques europees es desenvoluparan, s’elaboraran i difondran seguint aquest Codi.
A finals de setembre de 2011 el Comitè del Sistema Estadístic Europeu revisa la primera versió del codi. Es modifiquen lleugerament un dels principis i s’actualitzen els indicadors per avaluat-ne l’adopció efectiva.

Els 15 principis del Codi de Bones Pràctiques de les Estadístiques Europees s’agrupen en tres grups: Entorn institucional, procés estadístic i productes estadístics:

Entorn institucional

1. Independència professional
2. Mandat de recollida de dades
3. Adequació dels recursos
4. Compromís de qualitat
5. Confidencialitat estadística
6. Imparcialitat i objectivitat

Procés estadístic

7. Metodologia sòlida
8. Procediments estadístics adequats
9. Càrrega no excessiva per als enquestats
10. Relació cost/eficàcia

Productes estadístics

11. Rellevància
12. Precisió i fiabilitat
13. Oportunitat i puntualitat
14. Coherència i comparabilitat
15. Accessibilitat i claredat

Per a la temàtica del projecte són especialment rellevants els punts 5, 14 i 15.

3.1.1. Principi 5. Confidencialitat estadística

La privadesa dels informants (llars, empreses, administracions i altres enquestats), la confidencialitat de la informació que proporcionen i el seu ús exclusiu amb finalitats estadístiques estan totalment garantits.

Indicadors

5.1. La confidencialitat estadística està garantida per llei.
4. El sistema estadístico de l’Administració General de l’Estat

Fins a 1989 l’estadística espanyola està regida per la Llei d’Estadística de 1945 i el seu Reglament de 1948. En aquest període es van dictar importants regulacions de caràcter legal o reglamentari que van reforçar, actualitzar o complementar les disposicions originals, sobretot en aspectes relatius a centralització, coordinació, normalització i configuració del sistema estadístic estatal.

En arribar la transició política de mitjans dels anys setanta i el procés d’organització del nou Estat, es comença a plantear la necessitat de tenir una nova regulació de l’activitat estadística. L’any 1980 es presenta un Esborrany d’Avantprojecte de Llei d’Estadística, que pretén construir un sistema centralitzat, coordinat i dirigit per l’Institut Nacional d’Estadística, que finalment no va reeixir. En aquest sentit, la Constitució espanyola de 1978 atorga a l’Estat la competència exclusiva sobre
l’estadística amb finalitats estatales (Article 149.31), mentre que a la vegada els estatu
tuts d’autonomia de les comunitats autònomes naixents reconeixen la competència
estadística d’interès per a la comunitat autònoma.

A mitjans dels anys 80 es reprèn l’interès per elaborar una nova llei estatal d’es-
stadística. Mentre s’estava preparant l’Avantprojecte ja s’havien promulgat les lleis
d’estadística del País Basc (abril de 1986), Catalunya (juliol de 1987) i Galícia
(juliol de 1988), i altres Comunitats Autònomes tenien lleis similars en preparació.
Després d’un llarg procés de consulta a tots els Ministeris, Comunitats Autònomes i
diverses institucions el projecte s’aprova com a Llei de la Funció Estadística Pública
de 9 de maig de 1989. Els principals plantejaments de la Llei són:

- Regula exclusivament l’activitat estadística amb finalitats estatales portada a
terme per l’Administració General de l’Estat, bé sigui sola o en col·laboració.
  Deixa, per tant, fora del seu àmbit les estadístiques que es realitzen des de les
  Administracions Autonòmiques i Locals per les seves pròpies finalitats.
- Estableix quatre principis que ha de seguir l’elaboració de les estadístiques:
  secret, transparència, especialitat i proporcionalitat.
- Recull la uniformitat necessària dels marcs conceptuels per a les estadístiques
  que regula, especialment pel què fa a conceptes, definicions, unitats estadístiques,
  nomenclatures i codis que s’utilitzen.
- Determina que les relacions entre els serveis estadístics estatals i autonòmics
  s’estableixin mitjançant acords i convenis, ja sigui de forma bilateral o dins
  del Comitè Interterritorial d’Estadística, però amb l’obligació de facilitar-se
  les dades necessàries per a les estadístiques de respectiu interès, encara que
  aquestes dades estiguin sotmeses a secret estadístic.
- Impla la obligatorietat d’un Pla Estadístic Nacional, que al mateix temps
  és l’instrument per a definir les estadístiques amb finalitats estatals, i està
  acompanyat de Programes Estadístics Anuals que el desenvolupen. Tant el
  Pla Estadístic Nacional com els Programes Estadístics Anuals s’aproven per
  Reial Decret.
- Estipula que tots els serveis estadístics poden sol·licitar dades de les persones
  físiques i jurídiques, però que respondran de manera voluntària excepte si la
  corresponent estadística ha estat establerta per llei. Les dades més íntimes
  seran sempre de resposta voluntària.
- Regula el secret estadístic, sense recollir excepcions a la confidencialitat de les
  dades i protegint per igual tant a les dades estadístiques recollides directament
  com a aquelles que els serveis estadístics reben dels arxius administratius, però
  permet la circulació de dades confidencials entre diversos serveis estadístics.
- Estableix com a obligatòria la destrucció de les dades que serveixin per a la
  identificació directa dels informants, des del moment en què la seva conservació
  ja no sigui necessària. En tot cas, aquestes dades han de ser guardades sota
clos, precinte o dipòsits especials.
- No exigeix aprovació prèvia dels resultats abans de la seva difusió per part de
  cap òrgan diferent al que realitza la corresponent estadística.
- El Sistema Estadístic de l’Administració de l’Estat queda constituït per:
  - L’Institut Nacional d’Estadística (INE)
  - Les unitats estadístiques dels Departaments Ministerials i altres entitats
    dependents de l’Administració de l’Estat
5. El sistema estadístic català

Durant la transició espanyola, després del restabliment de la Generalitat l’any 1977, el govern de concentració nacional que forma el president provisional Josep Tarradellas publica en el segon Diari Oficial de 12 de gener de 1978 una Ordre que organitzava els Serveis de la Secretaria General de Presidència i creava el Servei Central d’Estadística i Documentació. Aquesta Ordre fa referència expressa al Servei Central d’Estadística de la Generalitat republicana que Lluís Companys crea l’any 1934. S’assignen al Servei quatre funcions: dirigir i coordinar la producció d’estadístiques de base, publicar els resultats, respondre consultes i intercanviar materials amb altres organismes similars.

L’octubre de 1978 un nou Decret del president Tarradellas va crear l’Institut Central d’Estadística i Documentació, adscrit al departament de Presidència i constituït pel Servei Central d’Estadística i pel Servei Central de Documentació. S’entenia que els diferents departaments podien tenir les seves pròpies oficines d’estadística, però s’esperava que l’Institut Central les coordinés i que també garantís la comparabilitat de resultats.

L’any 1979 s’aprova l’Estatut d’Autonomia de Catalunya, que en el seu article 9 disposa que Catalunya té competència exclusiva sobre estadística d’interés de la Generalitat.

El febrer de 1982 es determina que el Servei Central d’Estadística s’organitzi en quatre seccions: demografia, estadístiques econòmiques, mètodes i producció, i informació i difusió.

El setembre de 1984 es va suprimir l’Institut Central d’Estadística i Documentació, assumint les funcions estadístiques la Direcció General d’Afers Interdepartamentals. La decisió es basa en la duplicitat de funcions amb el Consorci d’Informació i Documentació de Catalunya. El Consorci era un organisme creat a principis dels anys setanta que fins aleshores pertanyia a l’àmbit local, emparat per les diputacions i les cambres de comerç, al quals s’incorpora la Generalitat a l’any 1983. El Consorci
es dedicava principalment al planejament urbà, la documentació i la producció estadística.

Tres anys més tard es promulga la Llei 14/1987, de 14 de juliol, d’estadística, ara ja derogada. En aquell moment és una llei d’una importància absoluta. Definía el Pla estadístic com a l’instrument d’ordenació de l’estadística de la Generalitat. Va introduir els principis d’objectivitat i correcció tècnica, l’obligació de col·laboració ciutadana i el seu recíproc, el secret estadístic. Va obligar a la difusió pública dels resultats estadístics i va preveure la creació de l’Institut d’Estadística de Catalunya així com les funcions que havia de dur a terme.


- El Parlament per mitjà de la Llei del Pla estadístic de Catalunya és qui fixa les estadístiques d’interès de la Generalitat, que tenen la consideració d’estadística oficial.
- Poden ser declarades d’interès de la Generalitat les estadístiques que fan:
  - l’Institut d’Estadística de Catalunya
  - els departaments de la Generalitat
  - les entitats de dret públic
  - els organismes autònoms i les empreses que en depenen
  - les entitats gestores de la Seguretat Social de Catalunya
  - els ens locals de Catalunya i els seus organismes autònoms i empreses que en depenen
  - les universitats, les cambres de Comerç, Indústria i Navegació i qualsevol altra entitat de dret públic si són sobre dades que cal obtenir a Catalunya o bé referides a Catalunya
  - l’Institut Nacional d’Estadística, els ministeris i les entitats de dret públic, els organismes autònoms i les empreses que en depenen, i altres organismes de l’Administració de l’Estat i de la Unió Europea si es fan sobre dades que cal obtenir a Catalunya o bé referides a Catalunya
  - instituts o centres de recerca universitaris, per encàrrec o bé mitjançant subvenció de les institucions o les entitats citades anteriorment si es fan sobre dades que cal obtenir a Catalunya i bé referides a Catalunya.
- Les institucions que formen part del Sistema Estadístic de Catalunya (SEC) són:
  - l’Institut d’Estadística de Catalunya, com a responsable de la coordinació i la gestió del SEC
els òrgans que tenen atribuïdes competències estadísticas dels departaments de la Generalitat, i dels organismes autònoms, les entitats de dret públic o les empreses que en depenen
els òrgans que tenen atribuïdes competències d’estadística oficial dels ens locals de Catalunya, i dels organismes, els ens o les empreses que en depenen
els òrgans o les dependències que fan activitats d’estadística oficial sobre dades que cal obtenir a Catalunya o bé referides a Catalunya de les universitats, les cambres de Comerç, Indústria i Navegació i les altres entitats de dret públic
els instituts i centres de recerca universitaris que fan activitats d’estadística oficial en el territori de Catalunya per encàrrec o bé mitjançant subvenció de les institucions o les entitats indicats en els punts anteriors

• Les estadístiques d’interès de la Generalitat:
  - Han de complir els criteris d’objectivitat i correcció tècnica:
    * Disposar d’un projecte tècnic que compleixi els requisits de la legislació vigent
    * Aplicar un sistema normalitzat de conceptes, definicions, classificacions i codis, i una metodologia que permeti la comparació amb els resultats d’altres estadístiques similars
    * Garantir una actualització periòdica
    * No duplicitat amb altres estadístiques existents
  - Preveure la màxima desagregació territorial de les dades
  - Tenir com a font prioritària d’informació els arxius i registres administratius i estadístics disponibles
  - Els resultats sintètics, que resumenixen de forma breu els resultats globals obtinguts, són de difusió obligatòria i accés gratuït a totes les persones i institucions
  - És obligatori donar informació, tant per part de les persones físiques com per les jurídiques. La informació sol·licitada s’ha de satisfer de manera completa i verífica, dins del període establert i respectant els formats i indicacions establerts
  - Complir estrictament la legalitat i en especial el secret estadístic.

• Preveu infraccions imputables tant als informants com al personal estadístic.

Al 2006 s’aprova el nou Estatut d’Autonomia de Catalunya. En el seu Article 135 sobre estadística es torna a indicar que la Generalitat té la competència exclusiva sobre estadística d’interès de la Generalitat, que inclou la planificació estadística, l’organització administrativa i la creació d’un sistema estadístic oficial propi de la Generalitat. La Generalitat participa i col·labora en l’elaboració d’estadístiques d’abast supraautonòmic.

L’any 2009 s’aprova un nou decret d’organització i funcionament de l’Institut d’Estadística de Catalunya (Idescat), que actualitza els òrgans de govern, l’organització i el funcionament de l’Idescat, adaptant-lo als nous reptes i objectius de l’estadística oficial catalana.

Capítol 2
Normativa aplicable a la protecció de dades

Donada la gran preocupació que genera la protecció del dret a la intimitat hi ha una extensa regulació al respecte, tant a nivell europeu com a nivell de l’Estat. En el cas d’Espanya aquesta regulació però, afecta només a les persones físiques.

1. Normativa relativa a la protecció de dades de caràcter personal

Podem separar aquest tipus de normativa segons si es tracta de normativa comunitària o normativa estatal. Tot i que la normativa comunitària s’ha de veure traslladada sempre en la normativa estatal pot passar un cert temps fins que no s’actualitzi i per tant és recomanable tenir-la en consideració.

No trobem normativa catalana al respecte ja que es tracta d’una competència de l’Estat central. Existeix l’Autoritat Catalana de Protecció de Dades, encarregada de supervisar únicament a les entitats del sector públic, però que aplica la normativa espanyola al respecte (Estatut d’Autonomia de 2006 en els seus articles 31 i 156).

1.1. Normativa europea

Els principals textos europeus en vigor relatius a la protecció de dades de caràcter personal són:

- Directiva 95/46/CE relativa a la protecció de les persones físiques en el que respecta al tractament de dades personals o a la lliure circulació d’aquestes dades (DOCE L 281 de 23.11.1995)
- Reglament 2011/45/CE relatiu a la protecció de les persones físiques en el que respecta al tractament de dades personals per les institucions i organismes comunitaris i la lliure circulació d’aquestes dades (DOCE L8 de 12.01.2001). Aquest reglament desenvolupa la Directiva 95/46/CE
2. NORMATIVA APLICABLE A LA PROTECCIÓ DE DADES

- Directiva 2002/58/CE relativa al tractament de dades personals i la protecció de la intimitat en el sector de les comunicacions electròniques (DOCE L 201 31.7.2002)
- Directiva 2006/24/CE sobre la conservació de dades generades o tractades en relació a la prestació de serveis de comunicacions electròniques d’accés públic o de xarxes públiques de comunicacions.
- Reglament 1367/2006 relatiu a l’aplicació, a les institucions i als organismes comunitaris, de les disposicions del Conveni d’Aarhus sobre l’accés a la informació, la participació del públic en la presa de decisions i l’accés a la justícia en matèria de medi ambient.

Actualment ja hi ha una proposta Directiva per substituir la Directiva 95/46/CE que està en una fase molt avançada de desenvolupament. Segons el que es menciona en les consideracions prèvies d’aquesta proposta de Directiva, la ràpida evolució tecnològica i la globalització han suposat nous reptes per a la protecció de les dades personals i aquests reptes requereixen l’establiment d’un marc més sòlid i coherent per la protecció de les dades a la Unió Europea que vingui recolzat per una execució estricta.

1.2. Normativa espanyola

Els principals textos espanyols en vigor relatius a la protecció de dades de caràcter personal són:

- La Constitució espanyola, en els seus articles 10, 14, 16, 18, 20, 53 i 105
- Llei orgànica 15/199 de protecció de dades de caràcter personal (BOE 298 de 14.12.1999). Correspon a la translació a la normativa espanyola de la Directiva 95/46/CE.
- Reial Decret 994/1999 pel qual s’aprova el Reglament de mesures de seguretat dels fitxers automatitzats que continguin dades de caràcter personal (BOE 151 de 25.06.1999)
- Reial Decret 1720/2007 pel qual s’aprova el Reglament de desplegament de la Llei orgànica 15/199 de protecció de dades de caràcter personal.

Les principals característiques d’aquests tres textos són:

- Dades especialment protegides:
  - Ningú pot ser obligat a declarar sobre la seva ideologia, religió o creences.
  - Les dades referents a l’origen racial, la salut i a la vida sexual només podran ser recaptades, tractades i cedides quan, per raons d’interès general, així ho disposi una llei o l’afectat ho consenteixi expressament.
  - Les dades de caràcter personal relatives a la comissió d’infraccions penals o administratives només poden ser inclòs en els fitxers de les Administracions públiques competents en els supòsits previstos en les respective normes reguladores.
- S’estableixen uns nivells de seguretat segons la naturalesa de la informació tractada, en relació a la major o menor necessitat de garantir la confidencialitat i integritat de la informació.
  - Nivell Bàsic
2. Normativa relativa al secret estadístic

La normativa comunitaria fa referències al secret estadístic però no proporciona cap pauta concreta sobre què queda afectat pel secret estadístic i com cal actuar. És per això que només cal centrar-se en la normativa espanyola, a partir de la Llei 12/1989 de la Funció Estadística Pública, i en la normativa catalana, a partir de la Llei 23/1998 d'Estadística de Catalunya.
2.1. El secret estadístic en la normativa espanyola

Podem trobar referència explícita al secret estadístic dins la normativa espanyola en la Llei 12/1989 de la Funció Estadística Pública on se li dedica per complet el Capítol III i es menciona en el Capítol VI sobre la difusió i conservació de la informació estadística.

Segons l’article 13 queden emparades per el secret estadístic les dades personals que els serveis estadístics obtinguin bé directament dels informants o bé a través de fonts administratives.

Es defineix per dades personals les referents a persones físiques o jurídiques que o bé permetin la identificació immediata dels interessats o bé conduexin per la seva estructura, contingut o grau a la identificació indirecta d’aquests. S’emfatitza que el secret estadístic no permet difondre en cap cas les dades personals sigui quin sigui el seu origen.

Aquesta definició és vàlida per a totes les Administracions i organismes públics.

Està prohibida la utilització de dades personals obtingudes directament dels informants per a finalitats diferents de les estadístiques (Article 14).

Està permesa la l’intercanvi de dades sotmeses a secret estadístic entre Administracions i organismes públics si:

- Desenvolupen funcions fonamentalment estadístiques
- Les dades són per a elaborar estadístiques que se’ls hi ha encomanat (estadística oficial)
- Disposin de mitjans per preservar el secret estadístic.

Excepcions al secret estadístic:

- Directoris que només continguin relacions d’establiments, empreses, explotacions o organismes de qualsevol tipus sempre que només continguin la seva denominació, ubicació, activitat i interval de grandària al que pertanyen. La dada de la grandària només es podrà mostrar si la unitat informant no manifesta expressament la seva disconformitat.
- En cas de consentiment dels afectats.

El personal estadístic i les persones físiques o jurídiques que, degut a la participació eventual en qualsevol fase del procés estadístic, tinguin coneixement de dades emparades pel secret estadístic han de preservar el secret estadístic fins i tot després de finalitzar les seves activitats professionals o la seva vinculació als serveis estadístics.

Una particularitat de la Llei és que les dades que serveixin per a la identificació immediata dels informants s’han de destruir quan la seva conservació ja no sigui necessària per al desenvolupament de les operacions estadístiques.

Durada del secret estadístic:

- Persones físiques:
2. NORMATIVA RELATIVA AL SECRET ESTADÍSTIC

- 25 anys des de la seva mort si la data és coneguda
- 50 anys a partir de l’obtenció de la dada

**Persones jurídiques:**
- Depèn de cada enquesta, però un temps no inferior a 15 anys

Les dades individuals no emparades pel secret estadístic es poden difondre sempre que sigui impossible identificar les unitats informants.

### 2.2. El secret estadístic en la normativa catalana

La normativa catalana respecte el secret estadístic es desenvolupa en la Llei 23/1998 d’estadística de Catalunya en el seu Capítol IV. En línies generals descriu el mateix que la normativa espanyola, tot i que també presenta algunes discrepàncies amb aquesta primera.

- No defineix què s’entén per dades de caràcter personal.
- Tot i que en l’Article 25 es diu que les dades obtingudes de fonts administratives estan subjectes al secret estadístic, en l’Article 28 se’n exclouen les dades obtingudes dels registres administratius, els quals poden tenir una normativa específica que els reguli i preservi.
- Les excepcions al secret estadístic són més àmplies que en el cas de la llei espanyola, i en alguns casos s’usen termes subjectes a la interpretació. Les excepcions no contemplades per la llei espanyola són:
  - Dades de coneixement públic i que no afectin a la intimitat de les persones.
  - Dades obtingudes de registres administratius
  - Respecte als directoris d’establiments, empreses, explotacions o organismes s’inclou “i altres característiques generals incloses habitualment en els registres o els directoris de difusió general”. No es fa cap referència a la possibilitat de negativa respecte a la dada de grandària.
  - Directoris d’edificis i d’habitatges que només contenen com a dades els identificadors, l’emplaçament, el tipus d’unitat i altres característiques generals que s’inclouen habitualment en els registres o els directoris de distribució general.
  - Es pot permetre a instituts d’investigació científica i als investigadors accedir a les dades emparades pel secret estadístic, sempre que aquestes dades no permetin una identificació directa de les persones i que aquestes institucions o persones compleixin les condicions adequades amb l’objecte de garantir la protecció física i informàtica de les dades emparades i evitar qualsevol risc de divulgació il·lícita.

En aquest sentit el Decret 143/2010 del Registre de Fitxers Estadístics i de les cessions de dades sotmeses a secret estadístic regula específicament com s’han de fer aquest tipus de cessions.

La Llei de la Funció Estadística Pública de l’Estat és més permissiva que la catalana (veure article 21.b)

- No es fa cap referència al temps que ha de passar fins que una dada personal pot fer-se pública.
Capítol 3
Nomenclatures i classificacions estadístiques

1. Característiques i conceptes clau

Segons la Divisió Estadística de les Nacions Unides es poden definir les classificacions estadístiques, altrament anomenades nomenclatures, classificacions de referència o classificacions estàndard com un conjunt discret, exhaustiu i mútuament exclonent d’observacions que poden ser assignades a una o més variables que es poden mesurar en una col·lecció de variants i/o presentació de dades.

L’estructura d’una classificació pot ser o bé jeràrquica o bé plana. Les classificacions jeràrquiques van des del nivell més ampli (per exemple divisió) fins al nivell més detallat (per exemple classe). Les classificacions planes (per exemple la classificació de gènere o sexe) no són jeràrquiques.

Les característiques d’una bona classificació són les següents:

- Les categories són exhaustives i mútuament exclonents (és a dir, cada membre d’una població només pot ser assignat a una categoria sense duplicitats o omissions).
- La classificació és comparable a altres classificacions estàndard (nacionals o internacionals).
- Les categories són estables (és a dir, no canvien amb massa freqüència, o sense l’apropiada revisió, justificació i documentació).
- Les categories estan ben descrites amb un títol amb format estàndard i acompanyades de notes explicatives, Índexs de codis, codificadors i taules de correspondència a les classificacions relacionades (inclusent versions anteriors de la mateixa classificació).
- Les categories estan ben balancejades entre els límits establerts per la classificació (és a dir, no massa categories o massa poques). Això normalment s’estableix aplicant criteris de significació.
- Les categories reflecteixen realitats del domini (per exemple la societat o l’economia) respecte al que estan relacionades (per exemple en una classificació
industrial les categories haurien de reflectir la imatge total de les activitats industrials del país).
- Disposa d’instruccions, manuals, índexs de codis i formació.

Les classificacions estadístiques s’usen per:
- Presentar informació estadística.
- Recopilar informació i/o organitzar informació ja recollida.
- Agregar i desagregar conjunts de dades amb finalitats d’anàlisi, incloent la construcció d’índexs.

Tot i que des de la UNECE s’equipara classificació estadística a nomenclatura tradicionalment s’ha distingit la primera de la segona per l’existència d’una estructura organitzada i ordenada dels seus termes. Segons això les classificacions jeràrquiques i les planes serien nomenclatures però només podríem usar el terme classificació per referir-nos a les classificacions jeràrquiques.

Les peces que conformen una classificació són les unitats estadístiques. Les unitats estadístiques són entitats, respondents d’una enquesta o coses usades amb la finalitat de calcular o mesurar. Les unitats estadístiques són unitats d’observació sobre les quals es recull o se’n deriva informació. Inclouen, entre d’altres, empreses, institucions governamentals, organitzacions d’individus, institucions, persones, grups, àrees geogràfiques i esdeveniments. Formen la població sobre la qual es poden recollir les dades o sobre les que es poden fer les observacions.

### 1.1. Famílies principals

Les classificacions estadístiques s’agrupen per tipologies temàtiques anomenades “famílies”, i les principals són les classificacions econòmiques (activitats econòmiques i productes) i les classificacions demogràfiques i socials (ocupacions, educació i sanitàries).

#### 1.1.1. Classificacions d’activitats econòmiques

Les classificacions d’activitats econòmiques serveixen per classificar les unitats productives (empreses i establiments) segons el tipus d’activitat a què es dediquin per elaborar estadístiques sobre els aspectes relacionats amb el seu funcionament econòmic: producció, utilització dels factors de producció (treball, capacitat, materials) o rendibilitat.

S’entén per activitat econòmica totaacció productora resultant d’una concurrència de mitjans o factors productius (equipament, mà d’obra, procediments de fabricació, productes, etc.) que porta a la creació d’un determinats béns o a la prestació de serveis concrets. Les activitats poden realitzar-se amb finalitats lucratives o sense.

Les classificacions d’activitats econòmiques s’agrupen en una única família que comprèn: CIIU, NACE, CNAE i CCAE.

La Clasificació industrial internacional uniforme per a totes les activitats econòmiques (CIIU) està elaborada per la Divisió Estadística de les Nacions Unides. La seva
darrera versió CIIU-Rev.4 és el resultat d’un important esforç de convergència i harmonització entre la Classificació industrial d’Amèrica del Nord (NAICS) i la versió anterior de la CIIU, la CIIU-Rev.3, perquè entre ambdues no quedava garantida la comparabilitat de les dades estadístiques.

A partir dels criteris fixats per la Divisió Estadística de les Nacions Unides Eurostat encapçalà la proposta els treballs de revisió de la Nomenclatura estadística d’activitats a la Unió Europea (NACE) vigent i a finals de l’any 2006 s’aprova el Reglament de la NACE-Rev.2, que esdevé de compliment obligatori a partir de 2008.

En base a aquest reglament l’Institut Nacional d’Estadística (INE) elabora la nova Classificació nacional d’activitats econòmiques 2009 (CNAE-2009), aprovada mitjançant Reial decret.


1.1.2. **Classificacions de productes**

Les classificacions de productes serveixen per classificar-los segons les necessitats de les estadístiques que les usen (producció, consum, comerç exterior, etc.). S’agrupen en tres principals famílies:

- Classificacions de comerç exterior: SH, NC i TARIC.
- Classificacions de productes per origen industrial: PRODCOM.
- Classificacions de béns i serveis: CPA, CNPA i CCPA.

Classificacions de béns i serveis: CPA, CNPA i CCPA.. Els governs i òrgans comercials necessiten disposar de dades exactes i comparables amb finalitats arancelàries i estadístiques. És per això que el Consell de Cooperació Duanera crea un Sistema harmonitzat per a la designació i codificació de mercaderies (SH) que són objecte de comerç internacional.

La Comunitat Europea implanta la Nomenclatura combinada (NC), que serveix tant per a referències fiscals com estadístiques. És una extensió del SH.

L’Aranzel integrat comunitari (TARIC) es construeix a partir d’una subdivisió complementària de la NC i té com a finalitat facilitar la informació dels complexos règims arancelaris aplicables per a la CE.

Classificacions de productes per origen industrial: PRODCOM. La classificació PRODCOM defineix l’estadística de producció de la Comunitat Europea per a les indústries extractives i les indústries manufactureres, així com per a la producció i distribució d’energia elèctrica, gas i aigua, és a dir, per a les seccions B, C i D de la NACE-Rev.2.
La base d’aquesta activitat estadística la constitueix un Reglament del Consell de les Comunitats Europees, relatiu a la creació d’una enquesta comunitària sobre la producció industrial (Reglament PRODCOM).

Classificacions de béns i serveis: CPA, CNPA i CCPA. Les classificacions de béns i serveis són un híbrid que sintetitza la informació referida als béns transportables, amb un criteri de classificació per origen industrial, i als béns no transportables o serveis que tenen com a objectiu recollir el resultat de les activitats de serveis.

Com a exemple tenim la Classificació Estadística de productes per activitats en la Comunitat Econòmica Europea 2008 (CPA-2008), que l’INE ha adoptat per Espanya i la Classificació catalana de productes per activitats 2014 (CCPA-2014) que en aquest moment elabora l’Idescat.

1.1.3. Classificació d’ocupacions

La classificació d’ocupacions té dos objectius bàsics: disposar d’un instrument estadístic de classificació laboral que permeti la comparabilitat, en l’espai i el temps, de dades estatals i internacionals; i disposar d’una llista uniforme de grups d’ocupacions que permeti als organizadors de programes de migració o de programes internacionals de formació professional identificar en els diferents països categories professionals comparables.

En aquestes classificacions les ocupacions s’agrupen en funció de la naturalesa del treball o la tasca realitzada. Els grans grups constitueixen amplis camps professionals. L’ocupació és l’últim nivell de treball, el més restringit, limitat i concret d’una classificació d’ocupacions. Aquest nivell engloba, en la seva denominació, ocupacions o càrrecs dels treballadors que realitzin qualsevol de les diferents combinacions que es puguin fer de les tasques.

L’Organització Internacional del Treball (OIT) és el responsable de la Classificació internacional uniforme d’ocupacions (CUIO). Aquesta classificació s’organitza d’acord amb el concepte de competència, definida com la capacitat de dur a terme les tasques inherents a un lloc de treball determinat, i deixa com a criteri subsidiari la relació de l’ocupació amb l’activitat. La versió actualment en vigor és la CUIO-2008.

La seva adaptació espanyola és la Classificació nacional d’ocupacions (CNO-2011) i l’adaptació al català d’aquesta darrera és la Classificació catalana d’ocupacions (CCO-2011), aprovada pel Decret 27/2012.

1.1.4. Altres classificacions bàsiques

Nomenclatures territorials (geonomenclatures). Les nomenclatures territorials (geonomenclatures) són un inventari de territoris, definits i delimitats, normalment ordenats segons un criteri (alfabetí o de situació geogràfica) que permeten la seva identificació d’una manera senzilla i inequívoca. La majoria són llistes simples de divisions territorials d’un total territorial considerat (per exemple la llista de comarques de Catalunya) i altres geonomenclatures estan estructurades en diferents nivells territorials relacionats jeràrquicament (per exemple la NUTS).
La finalitat d’aquestes classificacions és l’homogeneïtzació de la presentació de resultats estadístics per tal de garantir la comparabilitat espacial i temporal de la informació estadística associada al territori.

La International Standard Organization (ISO) té la responsabilitat d’establir la llista de països amb el seu sistema de codificació.

La Nomenclatura de les unitats territorials estadístiques (NUTS) és responsabilitat del Parlament Europeu i el Consell Europeu. La NUTS (del francès Nomenclature des Unités territoriales statistiques) és una nomenclatura geogràfica que subdivideix el territori de la Unió Europea en regions a tres nivells diferents (NUTS 1, 2 i 3 respectivament, des de les unitats territorials més grans fins a les més petites). Per sobre del nivell NUTS 1 tenim el nivell nacional de cada un dels Estats Membres. L’actual versió de la NUTS (2010) subdivideix el territori de la Unió Europea i els seus 27 Estats Membres en 97 regions NUTS 1, 270 regions NUTS 2 i 1294 regions NUTS 3.

Per a satisfer la demanda d’estadístiques a nivell local, Eurostat ha establert un sistema d’Unitats Administratives Locals (LAUs, de l’anglès Local Administrative Units) compatible amb la NUTS. El nivell superior de la LAU (LAU nivell 1, antigament NUTS nivell 4) està definit per a la majoria de països. Espanya no disposa d’unitats d’aquest nivell. El nivell inferior de la LAU (LAU nivell 2, antigament NUTS nivell 5) està formada pels municipis o unitats equivalents als 27 Estats Membres de la Unió. En total la Unió Europea està dividida en 120.419 unitats LAU 2.

Existeix també una geonomenclatura específica per a les estadístiques comunitàries relatives al comerç exterior amb tercers països. La versió actualment vigent està regulada pel Reglament 1106/2012 i està basada en la versió vigent de la norma ISO alfa-2.

En el cas de Catalunya, la conselleria d’Economia i Coneixement té la responsabilitat d’establir el sistema de codis territorials. Actualment està vigent l’Ordre 105/2005 pel qual s’estableix el sistema de codificar entre d’altres les comarques, els municipis, les províncies o les entitats municipals descentralitzades.

El nomenclàtor d’entitats de població és una nomenclatura que recull jeràrquicament les entitats inframunicipals, amb finalitats de recomptes de població.

En el cas de Catalunya les geonomenclatures contribueixen també en l’avanç cap a la normalització toponímica del català.


Les unitats estadístiques són els programes educatius i les qualificacions resultants. Les unitats estadístiques estan classificades en una jerarquia de nivells educatius basada en un nivell de complexitat creixent pel que fa a contingut educatiu. La CINE recull els programes educatius formals i no formals i qualsevol qualificació.
educativa formal reconeguda. La primera recollida de dades basada en aquesta nova classificació començarà a 2014.


L’objectiu bàsic de la CCED-2000 és el de constituir un instrument estadístic de classificació oficial que permeti un tractament homogeni de la informació derivada de l’àmbit de l’educació. Aquesta classificació es presenta des de tres èpiques que caracteritzen una mateixa realitat: la Classificació de programes per nivells educatius, la Classificació dels programes per sectors d’estudis i la Classificació de nivells de formació.

La classificació de programes per nivells educatius està estructurada de forma jeràrquica piramidal en dos nivells d’agregació.

La classificació de programes per sectors d’estudi estructura els programes educatius per àrees de coneixement o sectors econòmics als quals es refereix el contingut educatiu del programa.

La Classificació del nivell de formació és una aplicació de la Classificació de programes per nivells educatius, encara que aquesta recull els nivells de formació assolits.

Aquestes classificacions necessiten prendre en consideració els programes educatius més importants impartits en totes les èpoques, ja que es necessita tenir una visió històrica i dur a terme la comparabilitat entre programes educatius pertanyents a sistemes educatius de diferents períodes.

2. Obligatorietat d’ús de les classificacions

Al explicar les principals famílies de classificacions s’han anomenat diverses normatives al respecte. I és que moltes d’aquestes classificacions no impliquen un ús únicament estadístic sinó que tenen influència també en registres administratius o en els tràmits relacionats amb l’administració.

Per exemple, la Nomenclatura Combinada de productes s’usa per l’elaboració dels documents d’Intrastat que han de realitzar les empreses quan importen o exporten béns a països de la Unió Europea. Es tracta d’un tràmit administratiu amb finalitats únicament estadístiques que es va posar en marxa el Mercat Únic per a suplir les mancances d’informació que va suposar la desaparició de les formalitats duaneres. Aquest sistema de recollida d’informació està subjecte al Reglament (CE) 638/2004, al Reglament (CE) 1982/2004 i a la Resolució de 27 de gener de 2009 de la Presidència de l’Agència Estatal d’Administració Tributària. Al Reglament (CE) 1982/2004 en el seu Article 9 s’especifica que la informació de les mercaderies s’ha de fer seguint la Nomenclatura Combinada, establerta en el Reglament (CEE) 2658/87.
Tal i com s’ha anat veient anteriorment, les principals classificacions estadístiques venen acompanyades d’una normativa que les aprova. Això fa que trobem normativa europea, normativa espanyola (referent a l’adaptació al castellà i el cas particular de l’Estat espanyol si fos necessari) i normativa catalana (referent a l’adaptació al català i la realitat catalana al respecte si és necessari). Als llocs web d’Eurostat, l’Institut Nacional d’Estadística i l’Institut d’Estadística de Catalunya es pot trobar una llista detallada de tota la normativa relacionada amb les classificacions.

Per altra banda les operacions d’estadística oficial també tenen associada normativa europea al respecte. En moltes d’aquestes normatives es fa referència explícita a l’obligatorietat d’ús de determinades classificacions o a la obligatorietat de seguir determinats manuals on també consten classificacions d’obligat seguiment.

Per tant, com a norma general caldrà prendre en consideració les restriccions de les classificacions estadístiques a aplicar al tractar amb qualsevol operació estadística oficial, ja estiguin definides per una llei específica, un manual metodològic o el manual del productor de l’operació.
Capítol 4
Estadística de pensions de la Seguretat Social

Actualment les pensions es poden dividir en dos grans grups, segons siguin pensions contributives o no contributives.

Les pensions contributives (de la Seguretat Social) són prestacions econòmiques, generalment de durada indefinida, en les que la concessió està generalment supeditada a una relació jurídica prèvia amb la Seguretat Social (cal acreditar un període mínim de cotitzacions en determinats casos), sempre que es compleixin la resta de requisits. La seva quantia es determina en funció de les contribucions efectuades pel treballador o empresari, si es tracta de treballador per compte aliena, durant el període considerat a efectes de la base reguladora de la pensió que es tracti.

Les pensions no contributives són prestacions periòdiques o de duració determinada a favor de persones que no han cotitzat mai, o bé que no ho varen fer el temps suficient per aconseguir prestacions de nivell contributiu per a la realització d’activitats professionals. Aquesta classe de prestacions poden ser d’invalídesa o de vellesa.

A nivell competencial, l’Estatut d’Autonomia de Catalunya de 2006, en el seu Article 165 sobre Seguretat Social en el seu punt 1 apartat es reconeix com a competència compartida amb l’Estat el reconeixement i la gestió de les pensions no contributives. L’Estat manté per tant les competències normatives.

El fitxer que es tractarà en el projecte correspon només a les pensions contributives de la Seguretat Social, que és propietat de l’Institut Nacional de la Seguretat Social (INSS). Per tant s’expliquen només les característiques relatives a les pensions contributives.

Les pensions contributives de la Seguretat Social es presenten classificades per tipus i règim d’afiliació.

1. Tipus de pensions contributives

Les pensions contributives es separen en tres tipus:
Invalidesa o incapacitat permanent
• Jubilació
• Mort i supervivència

1.1. Pensions d’invalidesa o incapacitat permanent

La incapacitat permanent és una prestació econòmica de la Seguretat Social que, en la seva modalitat contributiva, cobreix la pèrdua de les rendes salarials o professionals que pateix una persona quan està afectada per un procés patològic o traumàtic derivat d’una malaltia o un accident i veu reduïda o anul·lada la seva capacitat laboral de forma presumiblement definitiva.

La incapacitat permanent es classifica en diversos graus, cadascun dels quals dona dret, si s’escau, a la corresponent prestació econòmica:

• Incapacitat permanent parcial per a la professió habitual
• Incapacitat permanent total per a la professió habitual
• Incapacitat permanent absoluta per a tot treball
• Gran invalidesa

En cas d’accident, es considera com a professió habitual l’exercida normalment pel treballador en el moment de patir-lo. En cas de malaltia, es considera com a professió habitual aquella a la qual el treballador dedicava la seva activitat professional abans de la incapacitat.

1.1.1. Incapacitat permanent parcial

És la que, sense arribar al grau de total, ocasiona al treballador una disminució no inferior al 33% en el seu rendiment normal per a la professió habitual, sense impedir-li fer-ne les tasques fonamentals.

Cal tenir menys de 65 anys en la data del fet causant i estar afiliat i en alta o en situació assimilada a l’alta quan es produeix el fet causant. Segons si la incapacitat deriva d’accident / malaltia professional o de malaltia comuna hi ha o no requisits mínims de cotització. La prestació s’abona en un únic pagament.
1.1.2. **Incapacitat permanent total**

És la que inhabilita al treballador a fer totes o les fonamentals tasques de la seva professió habitual, sempre que es pugui dedicar a una altra de diferent.

Com en el cas anterior, cal tenir menys de 65 anys en la data del fet causant i estar afiliat i en alta o en situació assimilada a l’alta quan es produeix el fet causant. Segons si la incapacitat deriva d’accident / malaltia professional o de malaltia comuna hi ha o no requisits mínims de cotització.

Es tracta d’una pensió vitalícia mensual que està en funció de la base reguladora i del percentatge aplicable, que excepcionalment es pot substituir per una indemnització a un tant alçat quan el beneficiari sigui menor de 60 anys.

És compatible amb el salari que es cobri per realitzar qualsevol treball per compte d’altres o pròpi, però incompatible amb l’exercici d’un treball corresponent a la professió que es va tenir en compte per a la declaració de la incapacitat.

Es pot extingir la prestació per:

- Revisió amb resultat de curació.
- Defunció de beneficiari.
- Reconeixement del dret de pensió de jubilació.
- Revisió d’ofi ci dictada per l’entitat gestora que resolgui la pèrdua del dret a la pensió.

1.1.3. **Incapacitat permanent absoluta**

És la que incapacitat completament el treballador per a tota professió o ofici.

Els requisits són molt similars als anteriors, però en aquest cas també es pot accedir a la pensió des de la situació de "no alta".

Es tracta d’una pensió vitalícia mensual que està en funció de la base reguladora.

És compatible amb l’exercici d’aquelles activitats, lucratives o no, compatibles amb l’estat de l’invàlid i que no representin un canvi en la seva capacitat de treball a efectes de revisió.

Es pot extingir per les mateixes causes que la incapacitat permanent total.

1.1.4. **Gran invalidesa**

És la situació del treballador afectat per una incapacitat permanent que, com a conseqüència de pèrdues anatòmiques o funcionals, necessita l’assistència d’una altra persona per als actes més essencials de la vida.

La prestació econòmica de gran invalidesa consta de dos elements: una pensió vitalícia corresponent a una incapacitat permanent total o a una incapacitat permanent absoluta més un complement perquè l’invàlid pugui remunerar la persona que l’atén.
Té els mateixos requisits i compatibilitats que la incapacitat permanent absoluta.

Les causes de suspensió i extinció són les mateixes que les exposades en la incapacitat permanent total.

1.1.5. *Incapacitat permanent - jubilació*

Les pensions d’incapacitat permanent passen a anomenar-se pensions de jubilació derivada d’incapacitat permanent quan els beneficiaris compleixen l’edat dels 65 anys.

1.2. *Jubilació*

La prestació econòmica per jubilació consisteix en una pensió vitalícia, única i imprescriptible que cobreix la pèrdua d’ingressos que pateix una persona quan, assolida l’edat establerta, deixa parcial o totalment la feina per compte propi o d’altri.

Les pensions de jubilació es classifiquen en quatre grups:

- Jubilació ordinària
- Jubilació anticipada
- Jubilació parcial
- Jubilació flexible

1.2.1. *Jubilació ordinària*

Cal complir dos requisits bàsics: edat i període mínim de cotització. La Llei 27/2001 sobre actualització, adequació i modernització del sistema de Seguretat Social fixa que l’edat de jubilació passa de manera gradual a ser de 65 anys a ser de 67 anys, amb un periode transitori comprès entre 2013 i 2017. El període mínim de cotització genèric és de 15 anys.

La quantia de la pensió es determina aplicant a la base reguladora el percentatge que correspongui en funció dels anys cotitzats. Existeixen supòsits de reducció i increment de les bases de cotització.

La pensió de jubilació és compatible amb els treballs per compte propi en els que es rebi uns ingressos anuals que no superin el salari mínim interprofessional, i amb les feines a temps parcial.

Aquesta pensió és incompatible entre d’altres supòsits amb la percepció de la pensió d’incapacitat permanent del mateix règim.

La pensió s’abona mensualment amb dues pagues extraordinàries al juny i al novembre. Està subjecta a IRPF i s’extingeix per la defunció del pensionista.
1.2.2. **Jubilació anticipada**

És la pensió iniciada abans de compliment de l’edat de jubilació ordinària, sempre que es compleixin determinats requisits i amb l’aplicació, si és el cas, de coeficients reductors.

En aquest grup cal destacar la jubilació anticipada per raons de grup o activitat professional desenvolupada. L’edat de jubilació ordinària que sigui d’aplicació, pot ser rebaixada o anticipada per la realització de determinades activitats professionals de natura excepcionalment penosa, perillosa, tòxica... Els col·lectius afectats per aquests coeficients reductors són les afectats per l’Estatut Miner, el personal de vol de treballs aers, els treballadors ferroviaris, els bombers al servei de les administracions i organismes públics, els artistes, els professionals taurins i els membres del cos de l’Ertzaintza.

El Bages i el Berguedà són dues comarques on l’activitat minera, tant pel que fa a l’extracció de potassa com a l’extracció de carbó, és molt important i on es d’esperar trobar un gran nombre de persones amb jubilació anticipada. Aquest fenomen serà més important en les localitats on es troben els centres de treball.

Un altre supòsit important en què es pot produir la jubilació anticipada és per raó de discapacitat.

Segons els supòsits pel qual s’accedeix a la jubilació anticipada es poden aplicar coeficients reductors de la pensió.

1.2.3. **Jubilació parcial**

És la iniciada després del compliment de l’edat marcada per llei (actualment de 61 anys), simultaneejada amb un contracte a temps parcial i vinculada o no amb un contracte de relleu.

Si el beneficiari és menor de l’edat de jubilació ordinària, l’empresa haurà de contractar simultàniament un contracte de relleu amb un treballador en situació d’atur.

Aquesta pensió és incompatible amb la pensió d’incapacitat permanent absoluta i gran invalidesa i amb la pensió de jubilació que pogués corresponder per una altra activitat diferent a la realitzada en el contracte parcial.

Els supòsits d’extinció de la pensió són la defunció del pensionista, el reconeixement de la jubilació ordinària o anticipada, el reconeixement d’una pensió d’incapacitat permanent declarada incompatible i l’extinció del contracte de treball a temps parcial excepte quan es tingui dret a l’atur o a altres prestacions substitutives.

1.2.4. **Jubilació flexible**

Es considera jubilació flexible la derivada de la possibilitat de compatibilitzar, un cop obtinguda, la pensió de jubilació amb un contracte a temps parcial.
1.2.5. **Règims especials i sistemes especials dins del règim general**

Els règims especials i els sistemes especials dins del règim general poden tenir particularitats normatives.

### 1.3. Prestacions de mort i supervivència

Són aquelles prestacions destinades a compensar la situació de necessitat econòmica que produeix, per a determinades persones, la defunció d’altres.

En aquest cas és el causant qui havia de tenir una relació prèvia amb la Seguretat Social. Els requeriments necessaris són diferents en cada un dels casos.

Les prestacions de mort i supervivència es classifiquen en 5 grups:

- Auxili per defunció.
- Pensió vitalícia de viduïtat.
- Prestació temporal de viduïtat.
- Pensió d’orfandat.
- Pensió vitalícia o subsidi temporal a favor de familiars.

En general aquestes prestacions estan subjectes a IRPF excepte les pensions derivades d’actes de terrorisme, les pensions d’orfandat i determinats tipus de pensions a favor de familiars.

#### 1.3.1. **Auxili per defunció**

El beneficiari és qui hagi tingut les despeses de l’enterrament. Es tracta d’un import fix.

#### 1.3.2. **Pensió de viduïtat**

A més dels requisits generals exigits al causant, els beneficiaris han d’acreditar uns altres requisits en determinades circumstàncies.

- El cònjuge supervivent. Si la defunció és derivada de malaltia comuna sobrevinguda abans del matrimoni cal complir un dels requisits següents: que el matrimoni s’hagi celebrat un any abans de la defunció, que hi hagi fills comuns, que en la data del matrimoni s’acrediti un període de convivència amb el causant que sumat a la durada del matrimoni superi els 2 anys.
- Les persones separades judicialment o divorciades. Quan tinguin assignada una pensió compensatòria.
- El supervivent amb el matrimoni declarat nul. Si se li ha reconegut el dret a la indemnització prevista a l’art. 98 del Codi Civil.
- El supervivent de la parella de fet. Defunció posterior a l’1 de gener de 2008. En el cas de Catalunya s’aplica en Dret Civil Català per acreditar la consideració de parella de fet.

La quantia de la pensió es determina aplicant un percentatge sobre la base reguladora. En els casos de separació judicial o divorci si hi ha concurrència de beneficiaris
amb dret a pensió, aquesta es reconeix en quantia proporcional al temps convis-cut per cadascun d’ells amb el causant garantit un mínim del 40% per al cònjuge supervivent.

La pensió de viduïtat és compatible amb les rendes del treball del beneficiari i amb la pensió de jubilació, vellesa i invalidesa SOVI o incapacitat permanent a què tingués dret, però amb aplicació de les regles de la concurrència i límits, mínims i màxims, de les pensions. La pensió de viduïtat a què pugui tenir dret l’orfe incapacitat és incompatible amb la pensió d’orfandat que estigui cobrant.

La pensió de viduïtat s’extingeix per tornar-se a casar o constituir-se com a parella de fet, d’acord amb la legislació específica, tot i que hi ha determinades excepcions. Tampoc es pot gaudir de la pensió si el beneficiari és declarat en sentència ferma culpable en la mort del causant. La pensió s’extingeix també per defunció del beneficiari.

1.3.3. Prestació temporal de viduïtat

És beneficiari el cònjuge supervivent quan no pot accedir a la pensió de viduïtat sempre que la defunció del causant derivi d’una malaltia comuna causada amb anterioritat al vincle matrimonial i que el beneficiari no pugui acreditar una de les tres situacions descrites anteriorment.

La quantia de la prestació és igual a la de la pensió de viduïtat que li hagués correspost, però amb una durada de 2 anys.

La prestació s’extingeix per el transcurs màxim de 2 anys o per les mateixes causes que la pensió de viduïtat.

1.3.4. Pensió d’orfandat

Els beneficiaris de la pensió d’orfandat poden ser els fills del causant o els fills que el cònjuge supervivent hagi aportat al matrimoni, sempre que la unió s’hagi celebrat 2 anys abans de la defunció del causant, hagi conviscut al seu càrrec i no tinguin dret a cap altra pensió de la Seguretat Social ni quedin familiars amb l’obligació i possibilitat de donar-los aliments, segons la legislació vigent.

En la data de la defunció del causant, els fills

- Han de ser menors de 21 anys
- Si són majors de 21 anys han de tenir reduïda la seva capacitat de treball en un percentatge valorat en un gran d’incapacitat permanent absoluta o gran invalidesa.
- L’edat de 21 anys es pot ampliar a 25 en determinats supòsits

La quantia es determina aplicant un percentatge sobre la base reguladora. Quan es tracti d’una orfandat absoluta les prestacions corresponents es poden incrementar.

La pensió d’orfandat és compatible amb les rendes del treball del cònjuge o parella de fet del causant, com amb la pensió de viduïtat que aquell percebi. Si l’orfe és menor de 21 anys o té reduïda la seva capacitat de treball en un percentatge valorat
en grau d’incapacitat permanent absoluta o gran invalidesa, la pensió s’abona amb independència dels ingressos derivats del seu treball.

La pensió d’orfandat que cobri l’orfe incapacitat que hagués contret matrimoni és incompatible amb la pensió de viduïtat a què posteriorment pogués tenir dret.

La pensió d’orfandat s’extingeix per complir l’edat màxima fixada en cada cas, per cessament de la incapacitat que va motivar la pensió, per adopció, per matrimoni (excepte si està afectat d’una incapacitat permanent absoluta o gran invalidesa) o per defunció de l’orfe.

1.3.5. *Prestacions a favor de familiars*

Es tracta de pensions o subsidis temporals atorgats a persones d’un altre parentiu que mantinguessin una relació de dependència amb la persona causant.

2. **Règims d’afiliació a la Seguretat Social**

El Sistema de la Seguretat Social és un conjunt de règims a través dels quals l’Estat garanteix a les persones compreses en el seu camp d’aplicació, per realitzar una activitat professional i també als familiars o assimilats que tinguin a càrrec seu, la protecció adequada en les situacions que la llei defineix.

Actualment el Sistema de la Seguretat Social està compost per:

- Règim especial de treballadors autònoms
- Règim General
- Règim especial de mineria del carbó
- Règim especial de treballadors del mar
- Assegurança escolar

A la Figura 4.2 es detaillen els sistemes especials col·lectius que actualment formen part del Règim General.

A la il·lustració següent es detallen els sistemes especials col·lectius que actualment formen part del Règim General.

Un cas particular és l’Assegurança Escolar. Dins d’aquest règim estan obligatòriament inclosos els estudiants espanyols o estrangers que resideixin legalment a l’Estat, que siguin menors de 28 anys, que estiguin matriculats oficialment a determinats plans d’estudi, que hagin abonat la quota corresponent i, per a determinats riscos coberts, que acreditin un període mínim de carència d’un any. Aquest règim dóna dret a pensions per incapacitat permanent, tot i que el nombre de casos és molt petit.

El règim d’afiliació segons el qual es classifica una pensió correspon al règim que va generar la pensió en el moment de sol·licituar-la, i per tant es poden trobar règims extingits o assimilats.
3. Recomanacions europees sobre estadística de pensions

Com a referència de les recomanacions europees sobre estadística de pensions, i prenent en consideració que les dades a tractar corresponen a l’any 2011, es prendrà com a referència el Manual Metodològic ESSPROS edició 2011 (”The European System of integrated Social PROtection Statistics””, el Sistema Europeu integrat d’Estadístiques de Protecció Social).

El Sistema ESSPROS va ser desenvolupat a finals dels anys 70 per Eurostat conjuntament amb representants dels Estats membres de la Unió Europea com a resposta a la necessitat de disposar d’un instrument d’observació estadística de la protecció social en els Estats membres.


Dos Reglaments de la Comissió sobre la recollida de dades van ser aprovats l’any 2011 (Reglament de la Comissió (UE) No 263/2011 i Reglament de la Comissió (UE) No 110/2011).

L’ESSPROS proporciona una comparació coherent entre els països europeus respecte a prestacions socials a llars i les seves finançaments. Les prestacions socials es transfereixen a les llars; bé mitjançant retribucions dineràries, bé en espècies; per aleujar la barrera financera de determinat nombre de riscos o necessitats.
Els riscos o necessitats de protecció social als quals l’ESSPROS fa referència es divideixen en diferents tipus: incapacitat, malaltia / atenció sanitària, vellesa, supervivència, família / fills, atur i exclusió social.

Les prestacions garantides dins del marc de la protecció social poden prendre diverses formes, tot i que el Manual les limita a

- Pagaments monetaris per protegir a la gent.
- Reembossaments de despeses efectuades per a la gent protegida.
- Béns i serveis proveïts directament a la gent protegida.

Les dades de les pensions de la Seguretat Social queden doncs inclòs dins de les retribucions dineràries en els camps de la incapacitat, la vellesa i la supervivència.

El Manual inclou una llista detallada de totes les taules a produir, ja siguin obligatòries o opcionals.

Destaca la diferenciació de les pensions de jubilació segons siguin ordinàries, anticipades o parcials. Aquesta distinció, tot i ser present en el sistema de pensions, no es veu reflectida en les dades.

El Manual ESSPROS s’organitza en diferents mòduls. Els mòduls que apliquen a les pensions de la Seguretat Social són el mòdul de beneficiaris de pensions i el mòdul de prestacions en protecció social.

3.1. Mòdul de beneficiaris de pensions

Segons el Parlament Europeu i la Regulació del Consell 458/2007 (Annex II), els beneficiaris de pensions es defineixen com els receptors d’una més retribucions dineràries periódiques d’un sistema de protecció social distribuïdes en set categories de pensions (pensió d’invalidesa, prestació de jubilació anticipada deguda a una capacitat reduïda per a treballar, pensió de jubilació, pensió de jubilació anticipada, pensió parcial, pensió de supervivència, prestació de jubilació anticipada deguda a raons del mercat laboral).

L’objectiu del mòdul de beneficiaris de pensions és calcular el nombre total de beneficiaris en cada una de les situacions següents:

- Cada una de les set categories de pensions.
- Cada una de les quatre categories que les agrupen (invalidesa, vellesa, supervivència i atur).
- A nivell global, com a agregació de les quatre categories anteriors.

En general, el nombre de pensions que reben els pensionistes difereix del nombre de pensionistes ja que molts pensionistes poden rebre més d’una pensió. El nombre total de beneficiaris es defineix per tant com el nombre de persones que reben com a mínim una pensió (és a dir, una persona que rep més d’una pensió només es compta una vegada). S’ha d’evitar el doble comptatge. Calcular el nombre de beneficiaris comporta l’agregació gradual al passar d’un nivell d’agregació a un altre.
Al Manual es detalla tot l’esquema de tabulació, tant l’obligatori com el voluntari. Per l’esquema obligatori cal separar les dades segons el gènere del beneficiari. La llista de definicions i classificacions dels beneficiaris incloses en la categoria d’obligatòries es descriuen al Reglament de la Comissió (CE) No 10/2008, Secció 2 Annexos I i II.

3.2. Mòdul de prestacions en protecció social

El Sistema ESSPROS recull dades de despesa en prestacions en protecció social a individus i llars sense la deducció de taxes o altres gravaments obligatoris sobre les prestacions en protecció social. Aquesta despesa és per tant considerada com a despesa bruta.

La comparabilitat entre països de les dades en despesa bruta està limitada pel fet que en alguns casos les prestacions pagades estan subjectes a determinats gravaments com taxes o contribucions socials mentre que en altres casos no ho estan. Les taxes o contribucions socials que graven les prestacions en protecció social reduixen la despesa bruta en protecció social dels governs i reduixen els ingressos dels que disposen els receptors.

Per tal de poder disposar de dades més comparables entre països sobre despeses en prestacions en protecció social es crea el mòdul de despeses netes en prestacions en protecció social, que és complementari al mòdul central de despeses brutes en prestacions en protecció social.

El càlcul de les despeses netes en prestacions en protecció social segueix la següent fórmula:

\[
\phi_n = \phi_b - \tau
\]

on:
- \(\phi_n\): Prestacions socials netes (restringits)
- \(\phi_b\): Prestacions socials brutes
- \(\tau\): Taxes i contribucions socials que paguen els beneficiaris

\[
\phi_n = \phi_b - \tau + B
\]

on:
- \(B\): Beneficis fiscals

entenent per benefici fiscals la protecció social concedida en forma de desgravacions que es defineixen com prestacions de protecció social si s’abonen en metàl·lic. Se n’exclouen les desgravacions per fomentar la concessió de prestacions socials o per fomentar els plans d’assegurances privades.
Capítol 5
La diferuició de resultats de l’Estadística de Pensions

El principal difusor de l’Estadística de Pensions és el Ministeri de Treball i Seguretat Social que és el ministeri del qual depèn l’Institut Nacional de la Seguretat Social. Aquest darrer, com ja hem vist anteriorment és el propietari de les dades corresponents a les pensions contributives de la Seguretat Social.

1. Ministeri de Treball i Seguretat Social

El Ministeri de Treball i Seguretat Social publica una informació relativament reduïda. La informació que es pot trobar al web del Ministeri es desglossa en les següents variables i categories:

Per tipus de pensió i règim de la Seguretat Social:

- Tipus de pensió:
  - Incapacitat permanent
  - Jubilació
  - Viduïtat
  - Orfandat
  - Favor de familiars

- Règims de la Seguretat Social:
  - Total
  - Règim General
  - Règim Especial de Mineria del Carbó
  - Règim Especial Agrari (compta aliena)
  - Règim Especial Agrari (compta pròpia)
  - Règim Especial de Treballadors Autònoms
  - Règim Especial de Treballadors de la Llar
  - Règim Especial de Treballadors del Mar
  - Accidents de treball
  - Malalties professionals
  - S.O.V.I.
• Any
• Mes

Per tipus de pensió i Comunitat Autònoma.

Per grau d’incapacitat permanent i règim de la Seguretat Social.

• Grau d’incapacitat permanent
  – Gran invalidesa
  – Incapacitat permanent absoluta
  – Incapacitat permanent total
  – Incapacitat permanent parcial

• Règims de la Seguretat Social:
  – Total
  – Règim General
  – Règim Especial de Mineria del Carbó
  – Règim Especial Agrari (compta aliena)
  – Règim Especial Agrari (compta pròpia)
  – Règim Especial de Treballadors Autònoms
  – Règim Especial de Treballadors de la Llar
  – Règim Especial de Treballadors del Mar
  – Accidents de treball
  – Malalties professionals
  – S.O.V.I.

• Any
• Mes

Per grau d’incapacitat permanent i règim de la Seguretat Social.

En una altra secció del web està disponible la informació de les pensions per tipus, règim de la Seguretat Social i Comunitat Autònoma i província només per el darrer mes.

2. Instituts d’Estadística de la resta de Comunitats Autònomes

La informació sobre pensions contributives de la Seguretat Social disponible als llocs web dels diferents Instituts d’Estadística de la resta de Comunitats Autònomes és molt diversa. En la taula següent es detalla la situació per a cada una de les Comunitats Autònomes:
<table>
<thead>
<tr>
<th>Comunitat Autònoma</th>
<th>Institut</th>
<th>Àrea temàtica</th>
<th>Informació disponible</th>
</tr>
</thead>
</table>
| Andalusia           | Institut d’Estadística i Cartografia d’Andalusia | Societat | • Despesa total en pensions segons tipus i sexe. Andalusia  
• Nombre de pensions en vigor i pensió mitjana per sexe. Andalusia i províncies. Dades anuals  
• Nombre de pensions en vigors segons règim, tipus i sexe. Andalusia i províncies. Dades anuals  
• Pensió mitjana per règim, tipus i sexe. Andalusia i províncies. Dades anuals  
• Nombre de pensions en vigor per grups d’edat, tipus i sexe. Andalusia i províncies. Dades anuals (edat en trams de 5 anys)  
• Pensió mitjana per grups d’edat, tipus i sexe. Andalusia i províncies (edat en trams de 5 anys) |
| Aragó               | Institut Aragonès d’Estadística | Mercat de treball | • Altes i baixes per tipus de pensions (Aragó i províncies) per mesos  
• Pensions en vigor per règim de la Seguretat Social (Aragó i províncies) per mesos  
• Beneficiaris i pensió mitjana per tipus de pensió (Aragó i províncies). Total anual |
| Astúries, Principat d’ | Institut Asturians d’Estadística | Estadístiques laborals | • Pensions segons règim de la Seguretat Social. Anual  
• Pensió mitjana segons règim de la Seguretat Social. Anual  
• Pensions màximes i mínimes segons tipus de pensió. Anual |
| Balears, Illes       | Institut d’Estadística de les Illes Balears. (IBESTAT) | Societat | • Altes i baixes de les pensions contributives per any  
• Nombre i import mitjà de les pensions contributives per any i tipus de pensió |
| Canàries            | Institut Canari d’Estadística (Istac) | Societat | • Tipus de pensió per mesos  
• Import mitjà per tipus de pensió per mesos |
| Cantàbria           | Institut Càntabre d’Estadística (ICA-NE) | Societat. Mercat de treball | • Tipus de pensió per mesos  
• Import mitjà per tipus de pensió per mesos  
• Altes i baixes per mesos |
<table>
<thead>
<tr>
<th>Comunitat Autònoma</th>
<th>Institut</th>
<th>Àrea temàtica</th>
<th>Informació disponible</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Castella i Lleó</td>
<td>Direcció General d’Estadística</td>
<td>Estadístiques socials</td>
<td>•Tipus de pensió per província i per mesos</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>•Import mitjà per tipus de pensió i província per mesos</td>
<td></td>
</tr>
<tr>
<td>Castella – la Manxa</td>
<td>Servei d’Estadística de Castella – La Manxa</td>
<td>Societat</td>
<td>•Tipus de pensió per província</td>
<td></td>
</tr>
<tr>
<td>Comunitat Valenciana</td>
<td>Institut Valencià d’Estadística (IVE)</td>
<td>Societat</td>
<td>•Tipus de pensió per província. Anual</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>•Segons Règim de la Seguretat Social. Anual</td>
<td></td>
</tr>
<tr>
<td>Extremadura</td>
<td>Institut d’Estadística d’Extremadura</td>
<td></td>
<td>No disposa de dades</td>
<td></td>
</tr>
<tr>
<td>Galícia</td>
<td>Institut Galleg d’Estadística (IGE)</td>
<td>Benestar i condicions de vida</td>
<td>•Nombre de pensions i import mitjà segons tipus de pensió i règim (Espanya, Galícia i províncies). Dades mensuals</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>•Nombre de pensions i import mitjà segons tipus de pensió i règim (Espanya, Galícia i províncies). Dades anuals</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>•Nombre de pensions i import mitjà segons gènere, edat i règim. Dades anuals (edat en trams de 5 anys)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>•Nombre de pensions i import mitjà segons gènere, edat i tipus de pensió. Dades anuals (edat en trams de 5 anys)</td>
<td></td>
</tr>
<tr>
<td>Madrid, Comunitat de</td>
<td>Institut d’Estadística de la Comunitat de Madrid</td>
<td></td>
<td>No disposa de dades</td>
<td></td>
</tr>
<tr>
<td>Múrcia, Regió de</td>
<td>Centre Regional d’Estadística de Múrcia (CREM)</td>
<td></td>
<td>No disposa de dades</td>
<td></td>
</tr>
<tr>
<td>Navarra, Comunitat Foral de</td>
<td>Institut d’Estadística de Navarra</td>
<td></td>
<td>No disposa de dades</td>
<td></td>
</tr>
</tbody>
</table>
3. Objectius de difusió

Per a la difusió de les dades de les pensions contributives de la Seguretat Social a Catalunya la es planteja un pla de tabulació més detallat que el corresponent al Ministeri de Treball i Seguretat Social i els de la resta de Comunitats Autònomes.

Es poden dividir els objectius de difusió en dos grans grups segons la grandària de la zona geogràfica: tabulació a nivell de Catalunya i òbits del pla territorial i tabulació a nivell municipal.

3.1. Tabulació a nivell de Catalunya i òbits del pla territorial

La llista dels creuaments de dades proposats és la següent:

- Nombre de pensions en vigor segons règim de la Seguretat Social, tipus de pensió i sexe. Catalunya i òbits territorials.
- Pensió mitjana en vigor segons règim de la Seguretat Social, tipus de pensió i sexe. Catalunya i òbits territorials.
5. LA DIFUSIÓ DE RESULTATS DE L’ESTADÍSTICA DE PENSIONS

- Import mensual total segons règim de la Seguretat Social, tipus de pensió i sexe. Catalunya i àmbits territorials.
- Nombre de pensions en vigor segons tipus de pensió, sexe i edat. Règim general. Catalunya i àmbits territorials. (edat en trams de 5 anys)
- Nombre de pensions en vigor segons tipus de pensió, sexe i edat. Altres règims. Catalunya i àmbits territorials. (edat en trams de 5 anys)
- Pensió mitjana segons tipus de pensió, sexe i edat. Règim general. Catalunya i àmbits territorials. (edat en trams de 5 anys)
- Pensió mitjana segons tipus de pensió, sexe i edat. Altres règims. Catalunya i àmbits territorials. (edat en trams de 5 anys)
- Nombre de pensions segons tipus de pensió, tram de percepció econòmica mensual i sexe. Règim general. Catalunya i àmbits territorials. (30 trams de percepció econòmica)
- Nombre de pensions segons tipus de pensió, tram de percepció econòmica mensual i sexe. Altres règims. Catalunya i àmbits territorials. (30 trams de percepció econòmica)

3.2. Tabulació a nivell municipal

A nivell municipal no hi ha una tabulació específica proposada. La intensió és poder donar el màxim d’informació possible per a tots els 947 municipis de Catalunya.

Es vol poder trobar un patró de publicació de resultats a nivell municipal segons la grandària dels municipis. La decisió de les taules a publicar en cada un dels conjunts de municipis així com la mateixa construcció d’aquests grups es prendrà en funció dels resultats obtinguts en la fase de protecció de les taules municipals.

![Figura 5.1. Estructura del pla de tabulació](image-url)
El llindar de 20.000 habitants s'inclou de manera automàtica ja que és el llindar que separa els municipis amb competències sobre serveis socials dels que no tenen aquestes competències.

L'estratègia a seguir per a poder establir els diferents nivells ha estat proposar una sèrie de taules, cada cop de complexitat major, observar el patró obtingut i relacionar els municipis sobre els quals no es pot donar informació amb la seva població. Amb aquesta informació es poden crear una sèrie de llindars relacionats amb les característiques dels municipis que ens permetin determinar diferents nivells d'informació.

Les taules proposades són:

- Nombre de pensions segons sexe.
- Nombre de pensions segons tipus (3 grans grups).
- Nombre de pensions segons tipus (3 grans grups) i sexe.
- Nombre de pensions segons règim (2 grups).
- Nombre de pensions segons règim (2 grups) i sexe.
- Nombre de pensions segons tipus (3 grans grups), règim (2 grups) i sexe.
Capítol 6
El Control de la Revelació Estadística

Al publicar informació estadística, els productors de la informació han d’aconseguir un equilibri entre els interessos de qui proporciona les dades i els interessos dels usuaris de les dades. Per una banda els usuaris estan interessats en obtenir una informació tan detallada com sigui possible per tal de rendibilitzar les despeses ocasionades per la recollida de dades. Per altra banda qui proporciona les dades (individus, empreses o altre tipus d’informants) requereixen que es garanteixi la seva privacitat i que es faci un bon ús de la seva informació. Tal i com ja s’ha vist en capítols anteriors no tan sols es tracta d’una qüestió de voluntat sinó que existeixen regulacions al respecte que obliguen els productors a garantir aquesta privacitat. En definitiva es tracta d’un equilibri entre dos drets: el dret a la informació i el dret a l’intimitat.

És fàcil veure que la millor manera d’assegurar la privadesa dels informants és no proporcionar cap tipus d’informació, mentre que publicar tota la informació suposaria no respectar la privadesa dels informants. Cal arribar doncs a aquest equilibri que permeti publicar dades i controlar la revelació estadística. Aquí, controlar la revelació estadística o SDC (Statistical Disclosure Control) significa evitar que es puguin extreure conclusions sobre unitats que es puguin reconèixer a partir de les dades que es fan accessibles al públic, és a dir, que es puguin identificar dades de caràcter individual vulnerant així la figura del secret estadístic.

Dins del procés de producció estadística, el control de la revelació acostuma a usar-se al final del procés: s’aplica immediatament abans de la publicació de les dades (sigui quin sigui el seu format). De manera ideal, s’hauria de prendre en consideració durant tot el procés de producció, ja que en darrera instància la publicació resultant ha de satisfer les polítiques de control de la revelació.

El control de la revelació estadística (veure definició de revelació estadística a la normativa catalana 2.2.2) es pot dividir en tres grans subgrups en què la metodologia ja està força ben definida.

- Mètodes de control de la revelació estadística per a microdades
- Mètodes de control de la revelació estadística per a dades tabulars
- Mètodes de control de la revelació estadística per l’anàlisi de resultats
Tot i així, al tractar-se d’una àrea de coneixement molt jove hi ha altres àmbits d’aplicació en els que també comença a desenvolupar-se, com és el cas de les aplicacions per a dades seqüencials o per a dades de mobilitat.

1. Mètodes de control de la revelació estadística per a microdades

Un conjunt de microdades $V$ es pot veure com un fitxer amb $n$ registres, on cada registre conté $m$ variables (o atributs) sobre un informant en concret (una persona, una llar, una empresa, etc.). Les microdades són la forma a partir de la que es deriven tota la resta de resultats i són la principal manera en què s’emmagatzemen les dades. Mentre que antigament els Instituts Nacionals d’Estadística només publicaven altre tipus de resultats (com per exemple dades tabulares), cada cop és més freqüent que les microdades esdevinguin una forma específica de difondre informació.

La finalitat del control de la revelació estadística en aquest cas és prevenir que es pugui inferir informació confidencial d’un informant concret a partir del fitxer de microdades. És a dir, donat un conjunt de microdades original $V$, l’objectiu del control de la revelació estadística és obtenir un conjunt de microdades protegit $V'$ de manera que:

- El risc de revelació (*id est*, el risc que un usuari o un intrús pugui usar $V'$ per determinar variables confidencials sobre un individu específic present a $V$) és baix.
1. MÈTODES DE CONTROL DE LA REVELACIÓ ESTADÍSTICA PER A MICRODADES 51

- Les anàlisis que puguin realitzar els usuaris sobre el fitxer \( V' \) (regressions, mitjanes, etc.) i sobre \( V \) donen lloc als mateixos resultats o com a mínim a resultats semblants.

1.1. Conceptes previs

1.1.1. \underline{Tipus de microdades disseminades}

Els conjunts de microdades protegides més habituals es coneixen amb els seus acrònims en anglès. Trobem els PUF (Public Use Files), o fitxers de microdades disponibles per a usuaris generals, i el SUF (Scientific Use Files), o fitxers de dades per a finalitats científiques. El nivell màxim de risc tolerable per un fitxer de tipus SUF serà superior al tolerable per un fitxer PUF.

![Gràfic Risc – Utilitat](image)

**Figura 6.2.** Gràfic Risc – Utilitat

1.1.2. \underline{Classificació de les variables d’un fitxer de microdades}

Depent de la seva sensibilitat, les variables d’un conjunt de microdades no protegides es poden classificar en quatre categories no necessàriament disjuntes:

- **Identificadors o identificadors directes:** Aquestes variables determinen de manera unívoca la identitat d’un informant, com per exemple el DNI, el NIF, el número de la Seguretat Social, etc. El control de la revelació estadística assumeix que els identificadors de \( V \) s’han eliminat o encriptat en un preprocés de les dades.

- **Quasi-identificadors o variables clau:** Un quasi-identificador és un conjunt de variables de \( V \) que, juntes, poden ser combinades amb informació externa per re-identificar (alguns) dels informants sobre els que \( V \) conté informació. Aquestes variables no poden ser eliminades, ja que potencialment qualsevol
variable pot formar part d’un quasi-identificador. En l’extrem s’haurien d’eliminar totes les variables per a assegurar que el conjunt de dades no conté cap quasi-identificador.

**Variables o atributs confidencials:** Variables que contenen informació confidencial o sensible sobre els informants

**Variables o atributs no confidencials:** Variables que contenen informació no sensible sobre l’informant. Les variables d’aquest tipus no poden ser obviades a l’hora de protegir un conjunt de dades, ja que poden formar part d’un quasi-identificador. En el cas de l’estadística oficial a Catalunya no podríem classificar cap variable com a no confidencial, ja que la legislació ens indica que totes les variables sense distinció han de ser protegides pel secret estadístic (2 2.2).

Depenent del tipus de dades que contenen, les variables d’un conjunt de microdades es poden classificar en:

- **Contínues:** Una variable es considera contínua si és numèrica i es poden efectuar operacions aritmètiques sobre ella. Alguns exemples poden ser l’edat o els ingressos.

- **Categòriques:** Una variable es considera categòrica quan pren valors sobre un conjunt finit i les operacions aritmètiques estàndards no tenen sentit. Els dos principals tipus de variables categòriques són:
  - **Ordinals:** Una variable ordinal pren valors en un conjunt ordenat de categories. Per tant, els operadors ≤, max, min tenen sentit. El nivell educatiu és un exemple de variable ordinal.
  - **Nominals:** Una variable nominal pren valors en un conjunt no ordenat de categories.

1.1.3. *Estimació del risc de re-identificació*

Actualment hi ha diferents tècniques per poder estimar el risc de re-identificació. Les principals tècniques són:

- Risc individual basat en la mostra: Regla del valor llindar, basada en les freqüències observades
- Estimació del risc individual a partir models que infereixen les freqüències poblacionals
  - usant pesos mostrals
  - ajustant models log-lineals
  - ajustant altres models
- Estimació del risc d’un registre mitjançant heurístiques
- Estimació mitjançant mètodes de record linkage

1.1.4. *Tipus de tècniques de protecció*

Actualment hi ha tres famílies principals de mètodes per a la protecció de microdades:

- Mètodes no pertorbatius
1. Mètodes de control de la revelació estadística per a microdades

- Mètodes pertorbatius
- Dades sintètiques i dades híbrides

1.2. Mètodes no pertorbatius

Els mètodes no pertorbatius no descansen en la distorsió de les dades originals sinó en la supressió parcial o la reducció del detall. Algun dels mètodes dins d’aquesta categoria poden ser aplicats tant a dades categòriques com a dades contínues, mentre que d’altres no són apropriats per dades contínues.

<table>
<thead>
<tr>
<th>Mètode</th>
<th>Dades contínues</th>
<th>Dades categòriques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mostreig</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Recodificació global</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Recodificació superior i inferior</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Supressió local</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

Taula 6.1. Mètodes no pertorbatius segons el tipus de dades

1.2.1. Mostreig

En lloc de publicar el fitxer original de microdades $V$ es publica una mostra $S$ del conjunt de registres originals.

El mostreig pot ser adequat per a variables categòriques, però en un escenari general és discutible la seva idoneïtat per a variables contínues. Suposem que es publica la variable $V_i$ sense modificar per a tots els registres de $S$. Aleshores, si la variable $V_i$ està present en un fitxer administratiu extern públic és molt probable que hi hagi correspondències úniques amb la mostra publicada.

1.2.2. Recodificació global

Per a una variable categòrica $V_i$ la recodificació global consisteix en combinar diverses categories per a formar-ne de noves i més generals, de tal manera que la variable resultat $V'_i$ té menys categories que la variables original $V_i$.

Per una variable contínua la recodificació global consisteix en substituir la variable $V_i$ per una variable $V''_i$ que és una versió discreta de $V_i$.

1.2.3. Recodificació superior i inferior (Top and bottom coding)

La codificació superior i inferior és un cas particular de la recodificació global que es pot fer servir en variables que presenten un ordre intern, és a dir, variables contínues o categòriques ordinals. La idea en la recodificació superior és agrupar els valors per sobre un cert llindar en una sola categoria. El mateix passa en el cas de la recodificació inferior.

1.2.4. Supressió local

Si una combinació de quasi-identificadors és compartida per molt pocs registres s’anomena una combinació no segura, ja que pot conduir a la re-identificació. En
aquest cas, alguns valors determinats de les variables que la conformen s’eliminen i es substitueixen per valors mancants.

1.3. Mètodes pertorbatius

Els mètodes pertorbatius d’emmascarament permiten publicar el conjunt sencer de microdades, tot i que els valors publicats són pertorbats en lloc d’exactes. Com en el cas anterior, no tots els mètodes pertorbatius són apropiats per tots els tipus de variables.

<table>
<thead>
<tr>
<th>Mètode</th>
<th>Dades contínues</th>
<th>Dades categòriques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emmascarament amb soroll</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Microagregació</td>
<td>x</td>
<td>(x)</td>
</tr>
<tr>
<td>Data swapping i Rank swapping</td>
<td>x</td>
<td>(x)</td>
</tr>
<tr>
<td>(intercanvi de dades)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data shuffling (barreig de dades)</td>
<td>x</td>
<td>(x)</td>
</tr>
<tr>
<td>Arrodoniment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remostreig</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>PRAM</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>MASSC</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

**Taula 6.2. Mètodes pertorbatius segons el tipus de variables. ’x’ significa aplicable i ’(x)’ significa només aplicable a variables categòriques ordinals**

La majoria de mètodes pertorbatius (com l’emmascarament amb soroll, el rank swapping o la microagregació) són casos especials d’emmascarament d’una matriu. Si la matriu original es $X$, aleshores les microdades emmascarades $Z$ es calculen fent

$$Z = AXB + C$$

on $A$ emmascara els registres, $B$ emmascara les variables i $C$ correspon a la matriu soroll.

1.3.1. **Emmascarament amb soroll (Noise masking)**

Es tracta del mètodes més senzill d’emmascarament. Correspondria a aplicar l’equació 3 amb les matrius $A$ i $B$ iguals a la identitat.

Hi ha dues grans famílies d’emmascarament amb soroll: mitjançant soroll additiu i mitjançant soroll multiplicatiu.

Els principals algorismes d’emmascarament amb soroll additiu presents a la literatura actual són:

- Emmascarament mitjançant soroll additiu sense correlació.
- Emmascarament mitjançant soroll additiu amb correlació.
• Emmascarament mitjançant soroll additiu i transformació lineal.
• Emmascarament mitjançant soroll additiu i transformació no lineal.

A la pràctica només s’aplica l’addició de soroll simple (en les seves primeres dues variants) o l’addició de soroll amb transformació lineal.

L’emmascarament amb soroll additiu és apropriat per variables contínues ja que:
• No fa cap suposició sobre el rang de valors possibles per a $V_i$.
• El soroll que s’afegeix és típicament continu i de mitjana zero.
• No és possible la correspondència exacta amb fitxers externs. Depenent de la quantitat de soroll que s’afegeixi pot ser possible trobar una correspondència aproximada.

Al projecte europeu CASC\(^1\) es va comprovar que els algorismes eren pràctics i usables, tot i que per la seva complexitat només poden ser aplicats per experts. Totes les aplicacions consumeixen molt de temps i requereixen coneixements experts tant de les dades com de l’algorisme.

Un dels principals problemes de l’emmascarament amb soroll additiu amb variància constant és que els valors petits són fortament pertorbats mentre que els valors grans no experimenten grans canvis. Els mètodes d’emmascarament amb soroll multiplicatiu volen suplir aquesta mancaça. Els dos principals tipus d’emmascarament amb soroll multiplicatiu són aquells que preserven els moments de primer i segon ordre i els que preserven diverses constriccions.

1.3.2. Microagregació

La microagregació és una família de tècniques de control de la revelació estadística orientades principalment al tractament de microdades contínues. La idea darrera de la microagregació és que les normes de confidencialitat permeten la publicació de conjunts de microdades si els registres corresponen a grups de $k$ o més individus on no hi ha cap individu que contribueixi excessivament al grup. L’aplicació estricta d’aquesta idea porta a substituir valors individuals amb valors calculats sobre petits agregats (microagregats).

Per a obtenir microagregats en un conjunt de microdades amb $n$ registres, aquests es combinen per formar $g$ grups de grandària com a mínim $k$. Per a cada variable, es calcula, per exemple, el valor mitjà en el grup i s’usa per substituir cada un dels registres originals. Els grups es construeixen usant un criteri de màxima similitud.

La microagregació existeix en diverses variants:

• Univariants vs. Multivariant
• Fixa vs. de grandària de grup variable
• Òptima vs. heurística
• Microagregació categòrica

\(^1\)Computational Aspects of Statistical Confidentiality, format per diferents Universitats i Instituts d’estadística i liderat per l’Institut Nacional d’Estadística d’Holanda CBS i que es va portar a terme entre els anys 2000 i 2003. http://neon.vb.cbs.nl/casc/CASCindex.htm
1.3.3. *Data swapping i rank swapping. Intercanvi de dades*

L’objectiu del data swapping és intercanviar el valor de variables confidencials entre registres individuals de manera que es mantinguin les freqüències de les categories amb recomptes petits o els marginals.

El rank swapping pot ser usat per qualsevol tipus de variable numèrica. Primer s’ordenen de manera ascendent els valors de la variable $X_i$, aleshores cada valor ordenat de $X_i$ s’intercanvia amb un altre valor ordenat de $X_i$ de manera que entre ells no hi hagi una diferència major del $p\%$ del total de registres.

1.3.4. *Data shuffling (barreig de dades)*

És un tipus especial de data swapping per a variables contínues i ordinals que garanteix que les distribucions marginals de les dades intercanviades és la mateixa que la de les dades originals. És un mètode sota patent dels Estats Units i per tant no pot ser implementat sense el consentiment dels autors.

1.3.5. *Arrodoniment*

Els mètodes d’arrodoniment substitueixen els valors originals de les variables per valors arrodonits. Aquests mètodes es poden aplicar de manera seqüencial (una variable cada vegada), amb la qual cosa tenim arrodoniment univariant, o també és possible fer un arrodoniment multivariant.

1.3.6. *Remostreig*

El mètode del remostreig es basa en extreure $t$ mostres independents de grandària $n$ (nombre de registres) de la variable original $X_i$. Aleshores s’ordena cada una de les mostres següent el mateix criteri. La variable emmascarada $Z_i$ es construeix de tal manera que l’element $x_{ij}$ és la mitjana dels $j$-èssims elements de les mostres ordenades $S_1, \ldots, S_t$.

1.3.7. *PRAM*

El mètode PRAM (Post Randomisation Method) és un mètode de control de la revelació estadística per a variables categòriques. El PRAM es pot considerar com una mala classificació intencionada, on les probabilitats de mala classificació són conegudes per qui protegeix les dades.

Quan s’usa aquest mètode, per cada registre del fitxer de microdades, el valor en un o més variables categòriques és canviat o no segons una certa probabilitat. Aquest procediment es realitza de manera independent per a cada registre. El mecanisme que determina les probabilitats de canvi es coneix per avançat i s’anomena matriu de transició o matriu de Markov.

Com que el PRAM és un mètode estocàstic el risc de revelació es veu directament afectat de manera positiva. A partir de les microdades protegides i la matriu de Markov es poden construir estimadors no esbiaixats de certs atributs estadístics de...
1. MÈTODES DE CONTROL DE LA REVELACIÓ ESTADÍSTICA PER A MIRCODADES

les dades originals. Per això l’usuari final ha de tenir per tant suficients coneixements estadístics per corregir els resultats del seu mètode d’anàlisi. Per aquest motivi els fitxers protegits amb aquest mètode només són recomanables per estadístics teòrics o amb àmplia experiència.

La selecció de les probabilitats de transició on és una tasca fàcil. No existeix una manera universal de prendre una decisió adequada per a cada situació.

Com que es tracta d’un mètode estocàstic el fitxer resultant després de d’una aplicació del mètode PRAM pot ser diferent, ja que aquest fitxer es tracta d’una realització d’un experiment probabilístic.

1.3.8. MASSC

El mètode MASSC es pot resumir en quatre etapes, que són les que donen lloc al seu acrònim:

1. Micro aglomeració per crear grups del conjunt de dades amb un risc de revelació semblant. Els estrats es creen usant quasi-identificadors.
2. Substitució probabilístètica òptima (com en el cas del mètode PRAM)
3. Submostreig (sub-sampling) per eliminar determinades variables o registres.
4. Calibratge dels pesos per preservar els estimadors de les variables modificades.

1.4. Dades sintètiques i dades híbrides

La publicació de dades sintètiques, és a dir simulades, és una alternativa a l’emmascarament per al control de la revelació estadística. La idea és generar aleatòriament dades amb la restricció que determinats estadístics o relacions internes de les dades originals es preservin.

Les dades sintètiques es poden usar com a mínim de tres maneres diferents:

- **Dades completament sintètiques**: No es publiquen dades originals, només dades sintètiques.
- **Dades parcialment sintètiques**: Només les dades més sensibles (és a dir algunes variables o registres) es substitueixen per valors sintètics abans de ser publicats. Per la resta, es publiquen els valors originals.
- **Dades híbrides**: Es combinen dades originals i dades sintètiques.

Les dades sintètiques són atractives en el sentit que, a primer cop d’ull, sembla que esquivin el problema de la reidentificació: com que els registres publicats són inventats i no es deriven de cap registre original es podria concloure que ningú es podria queixar d’haver estat reidentificat. Si per casualitat les variables no confidencials i les variables confidencials coincideixen amb les d’un ciutadà és fàcil que aquest percebi que les seves dades han estat revelades. En aquest cas és difícil que el ciutadà estigui content amb l’explicació que el registre va ser generat de manera sintètica, o fins i tot que entengui aquesta explicació.
Per altra banda les dades sintètiques tenen el problema de la seva limitada utilitat. Només es pot garantir que es preserven les propietats estadístiques que s’han pres en consideració explícitament en el model de generació de dades.

2. Mètodes de control de la revelació estadística per a dades tabulars

Les dades tabulares han estat tradicionalment el tipus de resultats que amb més freqüència han publicat els Instituts d’estadística. A partir de les microdades podem obtenir les dades tabulares creuant una o més variables categòriques. Tot i que les cel·les de les taules proporcionen dades agregades, també necessiten ésser protegides de possibles atacants.

2.1. Exemples de revelació en dades tabulars

2.1.1. El cas d’un atacant extern

Suposem que un Institut d’estatística publica les següents taules a nivell municipal. La taula 6.3 és una taula de freqüències de la variable “professió” segons “trams d’edat”. La taula 6.4 dona informació sobre el sou mitjà respecte les dues variables anteriors. És fàcil veure que estem donant la informació exacta del sou de l’únic tècnic o professional científic i intel·lectual del municipi.

<table>
<thead>
<tr>
<th>Ocupació</th>
<th>[16,24]</th>
<th>[25,35]</th>
<th>[36,45]</th>
<th>[46,65]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treballadors no</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qualificats</td>
<td>33</td>
<td>21</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Tècnics i professionals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>científics i intel·lectuals</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Treballadors qualificats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en activitats agràries i pesqueres</td>
<td>4</td>
<td>10</td>
<td>11</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Edat</th>
<th>[16,24]</th>
<th>[25,35]</th>
<th>[36,45]</th>
<th>[46,65]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treballadors no</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qualificats</td>
<td>33</td>
<td>21</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Tècnics i professionals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>científics i intel·lectuals</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Treballadors qualificats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en activitats agràries i pesqueres</td>
<td>4</td>
<td>10</td>
<td>11</td>
<td>6</td>
</tr>
</tbody>
</table>

Taula 6.3. Ocupació segons edat al municipi. Exemple: el cas del metge del poble

2.1.2. El cas d’un atacant intern

Si a la taula 6.3, en lloc de tenir un únic individu en la categoria de tècnics i professionals científics i intel·lectuals n’haguessim tingut dos, cada un dels individus podria haver descobert el sou de l’altre, ja que coneix la dada que ha proporcionat.

2.1.3. El cas d’un atacant intern amb dominància

Considerem el cas d’una gran empresa E que domina el seu sector empresarial, ja que representa, per exemple, el 90% del volum total de negoci del sector al territori. A partir de les dades principals del seu sector empresarial 6.5, i en combinació amb
2. SDC PER A DADES TABULARS

<table>
<thead>
<tr>
<th>Ocupació</th>
<th>Edat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treballadors no qualificats</td>
<td>16,24</td>
</tr>
<tr>
<td>Tècnics i professionals científics i intel·lectuals</td>
<td>2500</td>
</tr>
<tr>
<td>Treballadors qualificats en activitats agràries i pesqueres</td>
<td>1307</td>
</tr>
</tbody>
</table>

**TAULA 6.4. Mitjana del sou segons ocupació i edat al municipi.**
Exemple: el cas del metge del poble

les seves pròpies dades 6.6, l’empresa E podria obtenir una bona estimació de les dades de les seves competidores, i per tant es produeix una revelació indirecta de dades.

<table>
<thead>
<tr>
<th>Resultats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingressos d’explotació</td>
</tr>
<tr>
<td>Volum de negoci</td>
</tr>
<tr>
<td>Vendes netes de productes</td>
</tr>
<tr>
<td>Vendes netes de mercaderies</td>
</tr>
<tr>
<td>Prestacions de serveis</td>
</tr>
<tr>
<td>Altres ingressos</td>
</tr>
<tr>
<td>Despeses d’explotació</td>
</tr>
<tr>
<td>Consum de primeres matèries</td>
</tr>
<tr>
<td>Consum d’altres proveïmets</td>
</tr>
<tr>
<td>Consum de mercaderies</td>
</tr>
<tr>
<td>Treballs fets per altres empreses</td>
</tr>
<tr>
<td>Despeses de personal</td>
</tr>
<tr>
<td>Serveis exteriors</td>
</tr>
<tr>
<td>Inversions en actius materials</td>
</tr>
<tr>
<td>Resultat de l’exercici</td>
</tr>
</tbody>
</table>

**TAULA 6.5. Principals resultats. Exemple: el cas del gran grup empresarial**

2.2. Característiques de les dades tabulars

El nombre de registres \( r \) d’un fitxer de microdades és en general molt més gran que el nombre de cel·les \( n \) d’una taula \( (r \gg n \gg 0) \). És podria pensar doncs que la protecció d’un fitxer de microdades és molt més complexa, ja que suposa una gran quantitat d’informació. Malgrat això, les dades tabulars normalment tenen associades un conjunt de restriccions lineals \( m \). Aquestes restriccions lineals són les que modelitzen, per exemple, les relacions entre les cel·les internes d’una taula i les cel·les corresponents als totals.

La protecció de microdades requereix en general satisfer poques o cap constriccions lineals, mentre que en les dades tabulars el nombre de restriccions \( m \) acostuma a
### 6. CONTROL DE LA REVELACIÓ ESTADÍSTICA

| Ingressos d’explotació | 9232761.6 |
| Volum de negoci       | 8728493.4 |
| Vendes netes de productes | 8221686.7 |
| Vendes netes de mercaderies | 599533.9 |
| Prestacions de serveis | 145505.4 |
| Altres ingressos      | 383628.8 |
| Despeses d’explotació | 8981309.6 |
| Consum de primeres matèries | 3523326.3 |
| Consum d’altres proveïments | 2437412.0 |
| Consum de mercaderies | 428880.3 |
| Treballs fets per altres empreses | 17319.9 |
| Despeses de personal  | 812740.1 |
| Serveis exteriors     | 1103033.8 |
| Inversions en actius materials | 62012.2 |
| Resultat de l’exercici | 206182.4 |

**Taula 6.6. Principals resultats: Empresa E**

ser gran. És per aquesta raó que per a la protecció de dades tabulars es necessiten mètodes de programació lineal (LP) i mètodes de programació lineal entera mixta (MILP), la qual cosa fa que la protecció de taules grans i complexes sigui un problema difícil.

### 2.3. Tipus de taules

Les dades tabulars es poden classificar segons diversos criteris.

#### 2.3.1. Classificació segons el valor de les cel·les

- **Taules de freqüències o taules de contingència:** Contenen el nombre de registres que corresponen a cada una de les categories. Prenen valors en \( \mathbb{N} \).
- **Taules de magnituds:** Proporcionen informació sobre una tercera variable. Prenen valors en \( \mathbb{R} \).

#### 2.3.2. Classificació segons el signe de les cel·les

- **Taules positives:** Prenen valors \( \geq 0 \). És el cas més habitual.
- **Taules generals:** Els valors poden ser positius o negatius. Per exemple, les taxes de variació.

#### 2.3.3. Classificació segons l’estructura de la taula

- **Taula \( k \)-dimensional única:** Una única taula \( k \)-dimensial obtinguda a partir del creuament de \( k \) variables categòriques. El nombre de cel·les creix exponencialment a mesura que augmenta \( k \).
- **Taules jeràrquiques:** Conjunt de taules obtingudes a partir del creuament de diverses variables, algunes de les quals tenen una relació de jerarquia. Per
exemple, una variable jeràrquica és el codi CCAE. Un cas particular de taules jeràrquiques són les taules 1H2D, amb 1 variable jeràrquica i dues dimensions.

**Taules lligades (linked tables):** És el cas més general. Les taules lligades són un conjunt de taules obtingudes a partir d’un mateix fitxer de microdades. En teoria, tot el conjunt de taules obtingudes a partir d’un fitxer de microdades hauria de ser considerat a la vegada com un conjunt (més aviat molt gran) de taules lligades. El cas de les taules jeràrquiques és un cas particular de taules lligades. Per tant, la única manera segura de protegir totes les taules d’un fitxer de microdades seria protegir-les a la vegada com una única taula lligada.

2.4. Regles per detectar cel·les sensibles

2.4.1. **Regla de la freqüència mínima**

És la més usada per a taules de freqüències. En una taula de freqüències, una cel·la es considera sensibles si el nombre de contribuents a la cel·la és menor que $t$. Un valor freqüentment usat és $t = 3$. També es pot usar aquesta regla per a taules de magnituds, tot i que és preferible usar altres regles que tinguin en consideració les contribucions individuals a cada cel·la.

2.4.2. **Regla de concentració: dominància i $p\%$**

La regla de la dominància, també coneguda amb el nom de $(n,k)$ considera que una cel·la és sensible si $n$ o menys registres contribueixen en un $k\%$ (o més) del valor de la cel·la. La regla de dominància intenta evitar que una coalició de $n$ informants puguin obtenir una estimació acurada de la contribució de la resta d’informants. Alguns valors freqüents són per exemple ($n = 3$ i $k = 75$).

La regla del $p\%$ considera que una cel·la és sensible si un informant pot obtenir una informació de la contribució d’un altre informant amb una precisió del $p\%$. Normalment la regla del $p\%$ es prefereix a la regla de la dominància.

2.4.3. **Regla $(p,q)$**

És una extensió de la regla del $p\%$ que també s’anomena priori-posteriori. Abans de publicar la informació de la taula, el segon contribuent més gran a la cel·la té una informació a priori amb una precisió del $q\%$. A partir de la publicació de la taula el segon contribuent més gran a la taula podria extreure una informació del primer amb una precisió del $p\%$.

A partir d’aquesta regla una cel·la es considera sensible si, donada la variable $X$, que pren valors $x_1,\ldots,x_N$

$$X_R < \left( \frac{p}{q} \right) x_{(1)} , \quad \text{on} \quad X_R = \sum_{x=3}^{N} x_{(i)}$$

La informació d’un intrúss només pot augmentar si $p < q$. La regla $(p,q)$ és equivalent a la regla del $p\%$ si $q = 100$. 

2.4.4. *Regles de sensibilitat per casos especials*

**Contribucions negatives:** Es suggereix:
- Reduir $p$ en la regla del $(p\%$.
- Incrementar el valor de $k$ en la regla $(n,k)$ només respecte els valors positius.
- Usar la regla de la freqüència mínima.

**Renúncies:** Si un informant renuncia explícitament a la protecció de les seves dades s’han d’ajustar les regles de sensibilitat.

**Regla del “comerç exterior”:** En les estadístiques de comerç exterior tradicionalment s’ha aplicat una regla especial. Només per les empreses que de manera activa han demanat que es protegeixi la seva informació s’apliquen regles de sensibilitat \(^2\).

**Holdings empresarials:** En molts conjunts de dades, especialment els de caire econòmic, les unitats que representen els registres no corresponen amb les unitats que volem protegir. Sovint les empreses tenen diverses filials en diverses regions. Cal agrupar les contribucions d’una mateixa empresa abans d’aplicar les regles de sensibilitat.

**Pesos mostrals:** Si les microdades provenen d’una mostra també s’han de protegir les taules resultants. Per exemple podem tenir pesos mostrals iguals a 1. També té molt sentit en el cas d’enquestes amb errors mostrals petits i variables fortament esbiaixades. En aquests casos s’usen els pesos mostrals per determinar la sensibilitat de la cel·la.

2.5. Tècniques de protecció de dades tabulars

Degut a les relacions lineals existents entre les cel·les d’una taula, no és suficient protegir només les cel·les sensibles, a les que podem anomenar cel·les sensibles primàries, que es detecten amb l’aplicació de les regles de sensibilitat, sinó que es fa necessari modificar altres cel·les de la taula. Sense aquestes supressions addicionals es podrien recalcular els valors de les cel·les sensibles per diferenciació, usant les relacions lineals que existeixen entre les caseles d’una taula. Les cel·les addicionals que també s’han hagut de suprimir o modificar s’anomenen cel·les secundàries o complementàries.

Les tècniques de protecció de dades tabulars es poden classificar en dos grans grups: mètodes no-pertorbatius i mètodes pertorbatius.

<table>
<thead>
<tr>
<th>Mètodes</th>
<th>Pertorbatius</th>
<th>No pertorbatius</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recodificació de les variables</td>
<td>Controlled Rounding Problem</td>
<td>Cell Suppression Problem (CSP)</td>
</tr>
<tr>
<td>de creuament</td>
<td>Controlled Tabular Adjustment (CTA)</td>
<td></td>
</tr>
<tr>
<td>Protecció per intervals</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.5.1. Mètodes no-pertorbatius

Recodificació de les variables de creuament. *Global recoding*: Per a protegir les cel·les no segures de la taula el que es fa és recodificar les variables de creuament, de manera que diverses categories d’una variable es combinen per formar una nova categoria.

És tract d’una tècnica molt simple que no requereix de cap algorisme sofisticat. Aquesta estratègia modifica l’estructura de la taula i reduceix el detall de la informació, fent que les taules resultants a vegades pateixin una pèrdua d’informació molt gran.

És molt freqüent l’ús d’aquesta tècnica en els Instituts d’estadística, tot i que s’acostuma a combinar amb altres tècniques de protecció com la supressió de cel·les.

Cal recordar que la recodificació de variables s’ha de fer d’acord amb les classificacions estàndards o oficials.

Supressió de cel·les. *Cell suppression problem (CSP)*: En les taules protegides mitjançant aquest mètode, els valors que s’ha detectat que tenen risc de revelació són suprimits. A la pràctica això significa que són substituïts per un símbol especial. En el cas de l’Idescat aquest símbol acostuma a ser “:”.

Un procediment de supressió comporta dues fases:

- En la primera fase es detecten les cel·les sensibles primàries. Aquestes cel·les seran el que s’anomenen supressions primàries.
- En la segona fase, es seleccionen algunes cel·les no sensibles per protegir les cel·les sensibles primàries i impedir la revelació per diferenciació.

Aquestes conjunt de cel·les s’anomenen supressions secundàries.

El conjunt de supressions primàries i secundàries s’anomena patró de supressió.

<table>
<thead>
<tr>
<th>Taula original</th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>R₁</td>
<td>20</td>
<td>24</td>
<td>28</td>
<td>72</td>
</tr>
<tr>
<td>R₂</td>
<td>38</td>
<td>38</td>
<td>40</td>
<td>116</td>
</tr>
<tr>
<td>R₃</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td>TOTAL</td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taula protegida</th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>R₁</td>
<td>24</td>
<td></td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>R₂</td>
<td>38</td>
<td></td>
<td></td>
<td>116</td>
</tr>
<tr>
<td>R₃</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td>TOTAL</td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

Taula 6.8. Supressió de cel·les

Hi ha diferents mètodes de solucionar el problema de la supressió de cel·les:

- Mètodes d’optimització:
  - Mètodes d’optimització basats en programació lineal entera mixta (MILP) introduïts per Fischetti i Salazar al 2000 [*FS00*] basats en descomposició de Benders i plans de tall (detallat a [*FS01*]).
Mètodes d’optimització basats en una relaxació lineal del problema MILP implementats per Statistics Canada i descrits a [FFT09].

- Mètodes heurístics:
  - Solució mitjançant problema de camins mínims en xarxes per a taules 1H2D (1 variable jeràrquica i 2 dimensions) [Cas07].
  - Mètode de l’hipercub. Desenvolupat per l’oficina estadística de l’Estat del Rin del Nord Westfàlia a Alemanya³ per Repsilber i descrit a [GR02]. En aquest mètode, una taula jeràrquica es divideix en subtaules no jeràrquiques. Aleshores, les subtaules no jeràrquiques són protegides en un ordre determinat, on les subtaules en un nivell jeràrquic superior són tractades primer.
  - Mètode HiTas. Desenvolupat a l’Institut d’Estadística d’Hollanda (CBS) per De Wolf [De 02]. El mètode descompon les taules complexes en subtaules de 2 dimensions i les protegeix usant l’aproximació de programació entera mixta proposada per Fischetti i Salazar [FS00].

Protecció per intervals: En aquest cas les cel·les sensibles no són eliminades sinó que es publiquen els intervals de valors possibles per a cada una de les cel·les. És equivalent al resultat que obtindríem al voler reconstruir l’informació a partir de la taula amb les cel·les suprimides.

<table>
<thead>
<tr>
<th>Taula protegida</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>[0; 48]</td>
<td>24</td>
<td>[0; 48]</td>
<td>72</td>
</tr>
<tr>
<td>$R_2$</td>
<td>[10; 58]</td>
<td>38</td>
<td>[20; 68]</td>
<td>116</td>
</tr>
<tr>
<td>$R_3$</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td>TOTAL</td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

TAULA 6.9. Protecció per intervals

2.5.2. Mètodes pertorbatius

Existeixen dues grans famílies de mètodes pertorbatius: els aplicats abans de la realització de les taules i els aplicats un cop generada la taula.

Mètodes pre-tabulars: Soroll multiplicatiu: En aquest cas el que es proposa per protegir taules de magnituds és realitzar modificacions directament sobre les microdades. La versió més simple d’aquests mètodes consisteix en aplicar soroll multiplicatiu de mitjana zero i variància constant a les microdades.

L’avantatge d’aquests mètodes és que, un cop creat el conjunt de dades modificat, totes les taules que d’ell se n’extreguin satisfan dues condicions:
- L’additivitat de la taula.
- La consistència de les taules lligades.

³ Landesbetrieb Information und Technik NRW, abans anomenat Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen
Métode post-tabulars: Arrodoniment controlat: El mètode de l’arrodoniment controlat aconsegueix la protecció de les cel·les arrodonint totes les cel·les de la taula a un múltiple d’una determinada base \( r \). Els totals no poden ser arrodonits, ja que altrament no es compliria l’additivitat.

Un dels principals problemes del mètode és que força la desviació de tots els elements que no són múltiples de la base \( r \), la qual cosa reduceix la utilitat de la taula.

### Taula original

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 )</td>
<td>20</td>
<td>24</td>
<td>28</td>
<td>72</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>38</td>
<td>38</td>
<td>40</td>
<td>116</td>
</tr>
<tr>
<td>( R_3 )</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

### Taula protegida

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 )</td>
<td>25</td>
<td>24</td>
<td>30</td>
<td>75</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>120</td>
</tr>
<tr>
<td>( R_3 )</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>120</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>100</td>
<td>105</td>
<td>110</td>
<td>315</td>
</tr>
</tbody>
</table>

### Taula 6.10. Arrodoniment controlat

Métode post-tabulars: Ajust controlat (Controlled tabular adjustment, CTA):

L’objectiu de la metodologia de l’ajust controlat és trobar la taula additiva que s’assembli més a l’original garantint que els valors ajustats de totes les cel·les confidencials es troben prou lluny dels seus valors originals i que a la vegada els valors ajustats es troben dins d’un rang de valors determinat.

### Taula original

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 )</td>
<td>20</td>
<td>24</td>
<td>28</td>
<td>72</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>38</td>
<td>38</td>
<td>40</td>
<td>116</td>
</tr>
<tr>
<td>( R_3 )</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

### Taula protegida 1

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 )</td>
<td>15</td>
<td>24</td>
<td>33</td>
<td>72</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>43</td>
<td>38</td>
<td>35</td>
<td>116</td>
</tr>
<tr>
<td>( R_3 )</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

### Taula protegida 2

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 )</td>
<td>15</td>
<td>24</td>
<td>33</td>
<td>72</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>43</td>
<td>38</td>
<td>35</td>
<td>116</td>
</tr>
<tr>
<td>( R_3 )</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>121</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>98</td>
<td>101</td>
<td>110</td>
<td>309</td>
</tr>
</tbody>
</table>

### Taula 6.11. Ajust controlat

A la literatura actual es proposen diferents variants del CTA. Castro [Cas06] proposa solucionar el problema mitjançant la resolució d’un problema de programació lineal entera mixta. Com a alternativa de l’aproximació basada en
programació lineal Cox et al. [COS06] i Ichim i Franconi [IF06] proposen mètodes estadístics cpm l’“iterative proportional fitting” (IPF) o les tècniques de calibratge.

3. Mètodes de control de la revelació estadística per l’anàlisi de resultats

El problema de determinar si els resultats d’anàlisis estadístiques són prou segurs apareix bàsicament quan s’analitzen els resultats dels estudis a partir de microdades originals que es porten a terme en:

• els Centres d’Accés Segur per a investigadors als Instituts d’Estadística
• a través dels mitjans d’Accés Remot
• a través de l’Execució Remota

En principi aquest tipus de resultats també suposen un risc de revelació de dades d’informants individuals. El risc de revelació és present, per exemple, en el cas d’outliers.

Cal que els Instituts d’estadística comprovin els resultats dels anàlisis que els investigadors realitzen sobre dades no protegides abans que puguin fer-los servir o publicar-los. Respecte els resultats, no importa a través de quin dels mitjans anteriors s’hagin obtingut. En ambdós casos es realitzen el mateix tipus d’anàlisi amb el mateix tipus d’eines (R, SPSS, SAS, etc.).

Al tractar-se d’un problema comú a tots els Instituts que proporcionen aquest tipus de serveis la resolució d’aquesta classe de problemes es va desenvolupar a través d’un programa europeu finançat per Eurostat (Statistical Disclosure Control ESSnet project 2008-2009). Com a resultat d’aquest programa s’ha publicat una guia per a la comprovació de resultats (“Guidelines for Output Checking”). http://neon.vb.cbs.nl/casc/ESSnet/GuidelinesForOutputChecking_Dec2009.pdf

Es tracta d’una matèria que encara es troba en desenvolupament. Degut a la gran diversitat de casos que presenta el problema, tant en termes de possibles mètodes d’anàlisi com els diversos paquets estadístics, no ha estat possible desenvolupar un software específic.

Al comprovar els resultats d’una anàlisi estadística cal tenir en consideració dos possibles riscos:

• Risc d’autoritzar resultats amb risc de revelació.
• Risc de bloquejar resultats segurs.

A la guia es proposen dos tipus de mètodes:

• “A ull” (“Rule of Thumb”), que principalment minimitza el primer tipus d’error.
• Regla “Basada en principis”, que intenta minimitzar els dos errors.
La idea que hi ha darrera aquesta distinció és que la majoria dels resultats de recerca poden ser tractats amb una regla simple. Si aplicant la regla “A ull” un resultat resulta no permès però l’investigador vol que se li aprovi de totes maneres, aleshores cal fer treball extra (també per part de l’investigador) per demostrar que els resultats són segurs.

La llista del tipus de resultats que actualment es discuteixen a la guia és la següent:

<table>
<thead>
<tr>
<th>Estadístic descriptius</th>
<th>Taules de freqüències</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Taules de magnituds</td>
</tr>
<tr>
<td></td>
<td>Màxim, mínim i percentils (inclòs mediana)</td>
</tr>
<tr>
<td></td>
<td>Moda</td>
</tr>
<tr>
<td></td>
<td>Mitjanes, índexs, ratis, indicadors</td>
</tr>
<tr>
<td></td>
<td>Ratis de concentració</td>
</tr>
<tr>
<td></td>
<td>Altres moments d’ordre superior a 1 d’una distribució (inclouent variància, covariància, curtosi i apuntalament)</td>
</tr>
<tr>
<td></td>
<td>Gràfics: representacions pictòriques de les dades</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anàlisi de correlacions i regressions</th>
<th>Coeficients de regressions lineals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeficients de regressions no lineals</td>
</tr>
<tr>
<td></td>
<td>Estimació de residus</td>
</tr>
<tr>
<td></td>
<td>Resums i tests estadístic dels estimadors</td>
</tr>
<tr>
<td></td>
<td>Coeficients de correlació</td>
</tr>
</tbody>
</table>

Part 2

Resolució del cas
Capítol 7
Decisions preliminars

A l’hora de plantejar la resolució del problema, s’han hagut de prendre un seguit de decisions prèvies conjuntament per establir els criteris de protecció i els mètodes que calia aplicar.

1. Elecció de les dades sobre les que aplicar els mètodes

El fitxer de microdades sobre pensions de la Seguretat Social conté els registres de cada una de les pensions que han estat d’alta en algun moment del temps del període de referència, en aquest cas l’any 2011.

Com ja s’ha vist en la descripció del sistema de pensions de la Seguretat Social, existeix la possibilitat que un únic pensionista sigui beneficiari de més d’una pensió, dins del rang de compatibilitat de pensions existents. Això significa que qualsevol tabulació que es realitzi sobre dades de pensionistes no té per què complir la propietat additiva: el total de pensionistes d’un municipi no té perquè coincidir amb la suma del nombre de pensionistes per a cada tipus de pensions.

La teoria de control de la revelació estadística per a dades tabulars suposa sempre que les dades de què es disposa són additives (com a mínim en el nombre de contribuents a les caselles). Com que aquesta propietat no es compleix en les taules de pensionistes s’ha decidit treballar només amb les tabulacions corresponents a pensions.

Una altra decisió a prendre era la referència temporal de les dades. El fitxer correspon a una extracció de dades administratives amb data 31 de desembre. Tot i així, les dades disponibles no només inclouen les pensions en situació d’alta en aquesta data sinó que també s’inclouen els registres que estan en situació de baixa, bé sigui per defunció del beneficiari bé sigui per altres causes. En aquest sentit es pren la decisió de tabular només les pensions en situació d’alta en el moment d’extracció del fitxer.

1 veure 4.3.1
Per altra banda el fitxer també inclou informació de pensions generades a Catalunya, però on el beneficiari no resideix actualment a Catalunya. Com que determinats registres no disposen d’un número de la Seguretat Social vàlid per a poder corroborar aquesta informació i no suposen un nombre elevat de casos es desestima considerar-los en la tabulació.

L’univers de referència per a les tabulacions correspon per tant a les pensions de residents a Catalunya en situació d’alta a 31 de desembre de 2011.

2. Elecció dels mètodes de protecció

Una part important del procés de control de la revelació estadística és l’elecció dels mètodes a usar. La teoria només proporciona les eines per a dur a terme aquesta tasca, però són els propietaris de les dades qui han de prendre un paper actiu en la presa de decisions al respecte.

Com a regla per identificar cel·les sensibles s’escull el criteri més simple, el de la freqüència mínima, amb un valor de \( k = 3 \), ja que es considera que per la naturalesa de les dades no és necessari emprar tècniques més sofisticades.

Els instituts d’estadística han observat que el seu perfil mitjà d’usuari no se sent cóm que amb la idea de disposar de dades pertorbades, sigui quina sigui la magnitud de pertorbació. En canvi es pot acceptar més fàcilment la supressió d’informació. En base a aquesta premissa el mètode de protecció escollit és la supressió de cel·les (CSP).

3. Elecció del software

Actualment existeixen dues opcions de software disponibles per a la protecció de dades tabulars. Com a resultat del consorci establert entre diverses agències estadístiques de diferents països europeus es va desenvolupar una eina de control de la revelació estadística per a dades tabulars anomenada Tau-Argus que actualment és d’ús públic. Un altre software públic disponible és el programa sdcTable, desenvolupat per Bernhard Meindl de l’Institut d’Estadística d’Àustria com a un paquet estadístic per al software R.

Tot i que Tau-Argus és l’eina que es desenvolupà abans, a [SJE11] i [IF09] podem veure una llista prou acurada dels inconvenients que té aquest software:

- Per implementar els mètodes de control de la revelació estadística que usen Programació Lineal en Tau-Argus requereix de solvers externs i comercials, el preu dels quals no és gens menyspreable.
- Tau-Argus té el desavantatge comú en tots els softwares que tenen propietaris: no hi ha cap manera de comprovar, controlar, modificar o adaptar. El suport als usuaris només es pot obtenir a través de l’autor i actual encarregat de manteniment del programa.
- Només funciona sobre la plataforma Windows
4. PRE-PROCÉS DE LES MICRODADES

- Tau-Argus depèn d’un sistema de metadades complex que pot arribar a ser feixuc de fer servir.
- Tot i que disposa d’un manual, l’ajuda del software és pràcticament inexistènt. Alguns missatges d’error estan escrits en holandès.

Ichim i Franconi a [IF09] afirmen que un dels principals inconvenients d’R és que es tracta d’un software que no és fàcil d’usar i que per tant requereix una inversió inicial per aprendre a fer-lo servir, que vindria compensat per la seva gran flexibilitat. En aquest sentit cal apuntar, a títol personal, que aquest esforç inicial desapareix en el cas de disposar de plantilles joves de professionals estadístics, ja que és una eina d’ús habitual en la docència actual.

Tot i que sdcTable és relativament nou comparat amb Tau-Argus, els punts positius que els autor mencionats apunten sobre l’ús d’R són els següents:

- Els usuaris poden crear objectes per a qualsevol aplicació concreta.
- Hi ha diverses “mirror sites”, de manera que es garanteix la disponibilitat del software en tot moment.
- El software està disponible en diferents sistemes operatius (Windows, Linux i Macintosh), la qual cosa garanteix una gran difusió en la comunitat estadística.
- Al tractar-se de d’un software de codi obert els usuaris poden accedir al codi font per a qualsevol mètode estadístic emprat.

Considerant els avantatges potencials de fer servir el paquet d’R i els desavantatges que planteja l’ús de Tau-Argus s’ha pres la decisió de portar a terme la protecció de les dades tabulars usant sdcTable per a R.

4. Pre-procés de les microdades

Al tractar-se de microdades que encara no havien estat tractades va ser necessari fer un pre-procés inicial de les dades.

Per a dur a terme aquest tractament es va usar el software estadístic R. Es va crear un fitxer de lectura de les dades i codificació amb les característiques proporcionades per l’Institut Nacional de la Seguretat Social.

Una de les variables més importants del projecte correspon a la variable municipal. En realitzar un anàlisi descriptiu de les dades es va detectar que no es disposava de la variable codificada, sinó que la informació estava recollida de manera literal i amb un màxim de 150 caràcters. Va ser necessari usar una eina de codificació automàtica d’informació municipal de la qual disposa l’Idescat per a poder procedir a treballar amb les dades. Aquest fet va suposar un important cost en temps, ja que va caldre incloure aquest treball dins dels processos que ordinàriament es porten a terme.

Una estadística descriptiva de les dades va mostrar que hi havia petits problemes de depuració:

- La variable sexe no estava definida per a tots els registres del fitxer.
• No es disposava de la data de naixement per a tots els registres i per tant no es podia calcular la variable edat per a tots els registres.
• Hi ha algunes restriccions conceptuals que no es complien a les dades.
• No s'han pogut codificar els literals de tots els municipis.

Aquests casos s’han hagut de prendre en consideració al definir les estructures de les variables, ja que altrament no es compleix l’additivitat de les taules.

Les agregacions de categories en la variable règim de la pensió i les categories de les recodificacions de les variables contínues han estat proposades per la unitat responsable del tractament final de les dades i s’han executat els càlculs amb el software R.
Capítol 8
El problema de supressió de cel·les (CSP)

Tal i com ja es descriu en el Capítol 6.2, un mètode freqüentment usar per a protegir cel·les sensibles és la supressió (no publicar) determinades cel·les.

Aquest capítol es centra en detallar el problema de supressió de cel·les i els mètodes de resolució apropriats per a les taules que caldrà protegir.

Com que les taules que es vol generar no s’ajusten al patró de taules 1H2D (una variable jeràrquica i una taula de 2 dimensions) no es discutirà l’heurística de camins mínims en xarxes.

1. Definició del problema

La finalitat del problema de supressió de cel·les (CSP) és, donat un conjunt de cel·les sensibles al que anomenarem cel·les primàries, trobar un conjunt addicional de cel·les, al que anomenarem cel·les secundàries o complementàries, que garanteixi que el valor de les cel·les primàries no pot ser endevinat i que minimitzi determinats criteris de pèrdua d’informació.

En general podem definir el CSP com un problema de programació lineal entera mixta. Els paràmetres necessaris per caracteritzar el problema són els següents:

- Una taula general amb n cel·les i m relacions lineals $Aa = b$, on $a = (a_1 \ldots a_n)^T$ és el vector de valors de les cel·les.
- Fites inferiors i superiors $u$ i $l$ dels valors de les cel·les, que s’assumeixen coneguts per a qualsevol atacant, de tal manera que $l \leq a \leq u$ (per exemple, per una taula de valors positius $l = 0$ i $u = +\infty$).
- Un vector de pesos no negatius associats a la supressió de les cel·les $w_i$, $i = 1 \ldots n$. Per exemple, $w_i$ mesura el cost (o la utilitat de les dades) associades a la supressió de la cel·la $i$. La funció objectiu a minimitzar en el CSP és la suma dels pesos de les cel·les suprimides. Aleshores, si $w_i = 1$ el que minimitzem és el nombre total de cel·les suprimides.
8. EL PROBLEMA DE SUPRESSIÓ DE CEL·LES (CSP)

- Un conjunt $\mathcal{P} \subseteq \{1, \ldots, n\}$ de cel·les sensibles primàries, que s’ha obtingut prèviament aplicant alguna regla de sensibilitat.
- Llindars de protecció inferiors i superiors $lpl_p$ i $upl_p$ per a cada cel·la sensible $p \in \mathcal{P}$. Normalment aquest llindars es fixen com a un percentatge del valor de la cel·la $a_p$ o s’obtenen directament a partir de les regles de sensibilitat.

El CSP pretén trobar un conjunt $\mathcal{S}$ de cel·les secundàries a suprimir de tal manera que per a tot $p \in \mathcal{P}$:

\[
\begin{align*}
    a_p & \leq a_p - lpl_p & & i \quad \overline{a}_p \geq a_p + upl_p
\end{align*}
\]

on $a_p$ i $\overline{a}_p$ es defineixen com:

\[
\begin{align*}
    a_p = \min x_p & & \text{s.a. } Ax = b \\
    \text{s.a. } Ax = b & & l_i \leq x_i \leq u_i \quad i \in \mathcal{P} \cup \mathcal{S} \\
    x_i = a_i \quad i \notin \mathcal{P} \cup \mathcal{S} & & x_i = a_i \quad i \notin \mathcal{P} \cup \mathcal{S}
\end{align*}
\]

Kelly et. al (1992) són els primers en formular el model clàssic per al CSP. Es consideren dos conjunts de variables:

- $y_i \in \{0, 1\}, \ i = 1, \ldots, n$ La variable binària $y_i$ pren el valor 1 si la cel·la $i$ s’ha d’eliminar, altrament pren el valor 0.
- per a cada cel·la primària $p \in \mathcal{P}$, tenim dos vectors auxiliars $x^{l,p} \in \mathbb{R}^n$ i $x^{u,p} \in \mathbb{R}^n$, que representen les desviacions (positives o negatives) dels valors originals $a_i$. El model resultant és:

\[
\begin{align*}
    \min & \sum_{i=1}^{n} w_i y_i \\
    \text{s.a.} & \\
    Ax^{l,p} & = 0 \\
    (l_i - a_i)y_i & \leq x_i^{l,p} \leq (u_i - a_i)y_i \quad i = 1 \ldots n \\
    x_i^{l,p} & \leq -lpl_p \\
    Ax^{u,p} & = 0 \\
    (l_i - a_i)y_i & \leq x_i^{u,p} \leq (u_i - a_i)y_i \quad i = 1 \ldots n \\
    x_i^{u,p} & \geq upl_p \\
    y_i & \in \{0, 1\} \quad i = 1, \ldots, n
\end{align*}
\]

Les restriccions de desigualtat de 7 per la dreta i per l’esquerra imposen fites a $x_i^{l,p}$ i $x_i^{u,p}$ quan $y_i = 1$, i eviten desviacions en les cel·les que no es suprimeixen (és a
dir, si \( y_i = 0 \). Es pot veure que les restriccions imposades a 7 garanteixen que les solucions dels problemes lineals de 6 satisfaran les condicions requerides a 5.

El problema formulat a 7 és un problema de programació lineal entera mixta (MILP) amb:

- \( n \) variables binàries
- \( 2n|\mathcal{P}| \) variables contínues
- \( 2(m + 2n)|\mathcal{P}| \) restriccions

Es tracta d’un problema molt gran fins i tot per a taules d’una grandària moderada i amb poques cel·les sensibles. Es tracta doncs d’un problema NP-dur. Per exemple, per una taula amb 8000 cel·les, 800 cel·les sensibles i 4000 relacions lineals el MILP corresponent que obtindríem tindria 8000 variables binàries, 12.800.000 variables contínues i 32.000.000 restriccions.

Cal per tant estudiar altres aproximacions per poder resoldre aquest problema en casos reals.

### 2. Descomposició de Benders per al CSP

Es pot abordar el problema 7 mitjançant la tècnica de descomposició de Benders. Aquesta aproximació va ser inicialment plantegada per Fischetti i Salazar al 1999 [FS00] per a taules de dues dimensions. Posteriorment Fischetti i Salazar [FS01] van extreure-la a taules generals.

L’algorisme de Benders és un procés de generació de talls que resol de manera iterativa un problema mestre en les variables binàries \( y_i \in \{0,1\}, i = 1,\ldots,n \) – de manera que s’obté un patró de supressions – i \( |\mathcal{P}| \) subproblemes (un per a cada cel·la primària) que “informa” sobre el nivell de protecció que aquest patró proporciona per a cada cel·la primària. Si totes les cel·les primàries estan protegides, aleshores el patró de supressió és òptim. Altrament, s’afegeix al problema mestre un “tall de protecció” que no es compleix i es soluciona novament el problema mestre. Les iteracions es realitzen fins que es troba la solució òptima. L’obtenció de l’òptim està garantida ja que el nombre d’iteracions possibles és finit. A [Cas12] trobem la descripció detallada de l’aplicació de l’algorisme de descomposició de Benders:

1. **Inicialitzacions.** El conjunt inicial de restriccions de protecció és el conjunt buit \( \mathcal{J} = \emptyset \)

2. **Solucionar el problema mestre:**
8. EL PROBLEMA DE SUPRESSIÓ DE CEL·LES (CSP)

\[
\begin{align*}
\min & \quad \sum_{i=1}^{n} w_i y_i \\
\text{s.a.} & \quad y_p = 1 \, \forall p \in \mathcal{P} \\
& \quad y_i \in \{0, 1\} \, i = 1, \ldots, n \\
& \quad v_j^T y \geq \beta_j \, j \in \mathcal{J}
\end{align*}
\]

on \( v_j \in \mathbb{R}^n \) i \( \beta_j \in \mathbb{R} \) són els termes de l’esquerra i la dreta dels talls de proteCCIó (que es defineixen en les passes 3 i 4). Les cel··les primàries sempre es suprimeixen, fins i tot quan \( \mathcal{J} = \emptyset \).

3. Comprovar si el patró de supressions \( y_i, i = 1, \ldots, n \) satisfa els llindars de proteCCIó inferiors \( lpl_p \) per a cada cel··la primària \( p \in \mathcal{P} \):

- Per a cada cel··la primària \( p \in \mathcal{P} \) s’ha de comprovar si existeixen les desviacions \( x_{i,p}^{l} \) tal que satisfacin el primer grup de restriccions de 7

\[
\begin{align*}
Ax_{i}^{l,p} &= 0 \\
(l_i - a_i)y_i &\leq x_{i}^{l,p} \leq (u_i - a_i)y_i \quad i = 1 \ldots n \\
x_{i}^{l,p} &\leq -lpl_p
\end{align*}
\]

o de manera equivalent

\[
\begin{align*}
-lpl_p &\geq \min x_p \\
\text{s.a.} & \quad Ax_{i}^{l,p} = 0 \quad [\lambda] \\
x_{i}^{l,p} &\geq (l_i - a_i)y_i \quad i = 1, \ldots, n \quad [\mu_l] \\
x_{i}^{l,p} &\leq (u_i - a_i)y_i \quad i = 1, \ldots, n \quad [\mu_u]
\end{align*}
\]

on \( \lambda, \mu_l \) i \( \mu_u \) són el conjunt de multiplicadors de Lagrange (també coneguts com a variables duals) de cada grup de restriccions.

- El problema 9 sempre té solució ja que:
  - és factible. \( x_{i}^{l,p} = 0 \), és a dir cap desviació, és equivalent a la taula original i per tant és factible però no òptim.
  - no és un problema il··limitat, ja que \( x_p \geq l_p - a_p \geq -\infty \) (per exemple, si la taula és positiva aleshores \( l_p = 0 \)).

- Per dualitat en programació lineal, el problema dual de 9 es pot escriure com
2. DESCOMPOSICIÓ DE BENDERS PER AL CSP

\[
\text{max } 0\lambda + \sum_{i=1}^{n} (l_i - a_i)\mu_i y_i - \sum_{i=1}^{n} (u_i - a_i)\mu_i u_i = \\
= \sum_{i=1}^{n} ((l_i - a_i)\mu_i - (u_i - a_i)\mu_i u_i) y_i
\]

s.a. \[ A^T \lambda + \mu_l - \mu_u = e_p \]
\[
\mu_l \geq 0 \\
\mu_u \geq 0
\]

on \( e_p \) és la \( p \)-èssima columnà de la matrinx identitat.

- Aleshores, el nivell inferior de protecció de la cel·la \( p \) es satisfà si

\[
-l_p l_p \geq \sum_{i=1}^{n} ((l_i - a_i)\mu_i - (u_i - a_i)\mu_i u_i) y_i
\]

on \( \mu_l \) i \( \mu_u \) són la solució de 10

- Si es satisfà 11 per a tot \( p \in \mathcal{P} \), aleshores el patró de supressió y garanteix els límits de protecció inferior. Si, per a algun \( p \in \mathcal{P} \), no es satisfà 11, aleshores s’afegeix al conjunt de restriccions de protecció del problema mestre \( \mathcal{J} \).

4. Comprovar si el patró de supressió \( y_i, i = 1, \ldots, n \) satisfà els llimits de protecció superiors \( u_p l_p \) per a totes les cel·les \( p \in \mathcal{P} \):

- Es procedeix de la mateixa manera que en el cas dels nivells de protecció inferior. Si la protecció d’alguna cel·la sensible \( p \) no es compleix, aleshores s’afegeix una restricció de protecció al problema mestre (de manera similar que 11):

\[
u_p l_p \leq \sum_{i=1}^{n} (-(l_i - a_i)\mu_i + (u_i - a_i)\mu_i u_i) y_i
\]

on \( \mu_l \) i \( \mu_u \) són la solució del problema dual per a la protecció superior.

5. Si a les passes 3 i 4 no hi ha cap restricció que no es compleixi, aleshores el patró de supressió \( y_i \) és òptim. Altrament cal tornar a el pas 2 a resoldre el problema mestre amb el conjunt \( \mathcal{J} \) actualitzat.

Com en altres aplicacions de la descomposició de Benders, el temps d’execució del mètode depèn del nombre d’iteracions que calgui fer. En general, s’ha observat que aquesta aproximació és molt eficient per a la resolució del CSP en el cas de taules de dues dimensions. Desafortunadament, per a taules de grandària mitjana o gran i per a taules complexes, aquesta aproximació segueix essent computacionalment cara. És un mètode que garanteix l’obtenció de solucions òptimes si es disposa de suficient temps de CPU. A la pràctica, però, no sempre es disposa d’aquest temps.

És per això que cal estudiar altres alternatives per a la resolució del CSP.
3. El mètode de l’hipercub

Tal i com ja s’ha mencionat en el capítol 6.2 el mètode de l’hipercub és un mètode útil per a protegir taules jeràrquiques n-dimensionals.

3.1. Tractament de les taules lligades

Considerem com a exemple una taula de dimensió 3, $T_{1.3}$, definida per les següents variables: dimensió de l’empresa segons nombre de treballadors, codi CCAE i forma legal de l’empresa.

Si el nivell de detall és molt alt, la taula $T_{1.3}$ segurament no es podrà protegir en una única passa perquè el temps de càlcul seria massa gran. En aquesta situació es fa necessari descompondre la taula i protegir cada una de les seves parts.

La taula $T_{1.3}$ es pot descompondre en el següent conjunt de taules:

- $T_1$: empreses segons nombre de treballadors.
- $T_{1.1}$: empreses segons nombre de treballadors i codi CCAE.
- $T_{1.2}$: empreses segons nombre de treballadors i forma legal.

Aleshores, $T_1$ és una subtaula de $T_{1.1}$ i també de $T_{1.2}$. Una cel·la serà sensible a $T_1$ si i només si ho és a $T_{1.1}$ i a $T_{1.2}$ individualment.

Si realitzem una supressió de cel·les secundàries a $T_{1.1}$ i a $T_{1.2}$ de manera separada és molt probable que hi hagi cel·les de $T_1$ que no se suprimeixin a $T_{1.1}$ que sí es suprimeixin a $T_{1.2}$ i a l’inrevés.

El mètode de l’hipercub es basa doncs en protegir cada una de les taules de manera individual i a la vegada guardar la informació relacionada amb les noves supressions de les cel·les compartides entre les diverses taules.

En el cas de l’exemple l’algorisme funcionaria de la següent manera:

1. Solucionar el CSP per a la taula $T_{1.1}$. Guardar la informació de les supressions secundàries a les cel·les que $T_{1.1}$ i $T_{1.2}$ comparteixen, és a dir, $T_1$.

2. Les supressions secundàries a $T_1$, resultants de la protecció de $T_{1.1}$, es consideren com a supressions primàries a $T_{1.2}$. Solucionar el CSP per a $T_{1.2}$.

3. Si hi ha hagut noves supressions secundàries a $T_1$, considerar les cel·les suprimides com a supressions primàries a $T_{1.1}$.

4. Repetir les passes 1–3 fins que no hi hagi noves supressions secundàries a $T_1$.

Un cop finalitzat el procediment, qualsevol cel·la compartida entre $T_{1.1}$ i $T_{1.2}$ (cel·les de $T_1$) o bé estarà suprimida tant a $T_{1.1}$ com a $T_{1.2}$ o bé no estarà suprimida en cap taula. A més a més, cap de les supressions pot conduir a la revelació d’informació fent servir les relacions lineals existents entre les cel·les suprimides i no suprimides de $T_{1.1}$ o $T_{1.2}$, ja que la solució del CSP en els passos 1 i 2 ho garanteix.
3.2. Algorisme de resolució del CSP

Aquest algorisme es basa en el fet que no es pot revelar informació d'una cel·la suprimida en una taula simple \( n \)-dimensional sense jerarquia si aquesta cel·la forma part d’un patró de supressió format per cel·les de valor diferent a zero i que corresponen a les cantonades d’un hipercub\(^1\).

1. Les taules \( n \)-dimensionals amb estructura jeràrquica es subdivideixen en un conjunt de subtaules \( n \)-dimensionals sense estructura.

2. Comencem tractant la subtaula de nivell jeràrquic superior.

3. Per a cada supressió primària de la subtaula que s’està tractant, es construeixen tots els hipercubs on aquesta cel·la és una de les cantonades.

4. Per a cada hipercub, es calcula una fita inferior de l’amplada de l’interval de supressió per a la cel·la sensible. Aquesta fita es calcula a partir del que suposaria suprimir les cel·les situades a les cantonades de l’hipercub. No és necessari resoldre cap problema de Programació Lineal per a trobar aquesta fita. Si la fita és suficientment ampla, l’hipercub es considera factible.

5. Per a cada un dels hipercubs factibles, es calcula el valor de la pèrduda d’informació associada a suprimir les cel·les que en conformen les cantonades. Es selecciona l’hipercub amb menor pèrduda d’informació i es suprimeixen les cel·les corresponents a les cantonades.

6. Seguint l’ordre jeràrquic, es realitzen els passos 3–5 per a totes les subtaules.

7. Un cop s’han tractat totes les subtaules, totes les cel·les que s’han suprimit de manera secundària i que pertanyen a més d’una subtaula es consideren com a supressions primàries. Repetir les passes de 2 a 6.

Aquest mètode proporciona un criteri suficient però no necessari per a l’obtenció d’un patró de supressió. És a dir, per a un cas particular de subtaules el “millor” patró de supressió pot ser que no sigui l’obtingut amb l’algorisme de l’hipercub, la qual cosa suposa que la solució obtinguda sobreprotegiria la taula.

L’algorisme incorpora una variant per a protegir el cas en què un únic informant contribueixi a una cel·la. En aquest cas, per assegurar que aquest informant no fa servir la seva informació per a recalcular els valors de les restes de cel·les suprimides en l’hipercub, s’imposa que la cel·la sigui a la cantonada de com a mínim dos hipercubs.

Com és ben sabut, a partir d’una taula de valors no negatius amb un determinat patró de supressió és possible aconseguir fites inferiors i superiors als valors de les cel·les eliminades. Aquest problema és l’anomenat “problema d’auditoria” o auditing. L’interval que s’obté a partir d’aquestes fites s’anomena “interval de supressió”. Les implementacions d’aquest algorisme demanen que l’usuari proporcioni un rang mínim de l’interval de supressió, normalment com a una ràtio del

\(^1\)Veure [Rep95]
valor total de la cel·la. Un hipercub es considera factible si l’amplada de l’interval de protecció és superior al rang mínim que l’usuari ha establert.

3.2.1. Càlcul de la pèrdua d’informació

La mesura de pèrdua d’informació que es fa servir és normalment proporcional al logaritme del valor de la cel·la. Aquesta mesura s’acostuma a usar en combinació amb altres aproximacions heurístiques per a diferenciar entre diferents categories de cel·les, com les cel·les suprimides, les no suprimides, les cel·les amb valor zero i les cel·les amb un únic contribuent. El cost de suprimir un hipercub que contingui com a mínim una cel·la no suprimida serà major que el cost de suprimir un hipercub que només contingui cel·les ja suprimides.

Hi ha també la possibilitat de que l’algorisme eviti suprimir cel·les als marginals de les subtaules. Amb aquesta opció, a les cel·les dels marginals d’una subtaula se’ls hi assignen uns costos addicionals alts. Es recomana aplicar aquesta opció, no només perquè les cel·les dels marginals contenen més informació, sinó perquè són aquestes cel·les les que creen els lligams amb les altres subtaules d’una taula complexa. Qualsevol cel·la que es suprimeixi en el marginal d’una subtaula pot produir supressions addicionals en una altra subtaula.

Hi ha una altra possibilitat que permet fer quelcom semblant a una “recodificació local”. Amb aquesta opció les cel·les adjacents d’una classificació de grandària tenen preferència a l’hora de ser triades com a supressions secundàries. En determinats casos això permet “recodificar” localment la classificació de grandària.

4. El mètode HiTas

Tal com ja s’ha mencionat a 6.2, el mètode HiTas és una heurística per a resoldre el problema de la supressió de cel·les en taules jeràrquiques.

De manera esquemàtica, el mètode HiTas es defineix de la següent manera:

1. Dividir la taula jeràrquica en totes les subtaules no jeràrquiques que tinguin sentit.

2. Agrupar les subtaules en classes de manera que totes les taules d’una mateixa classe es puguin protegir de manera independent.

3. Protegir totes les taules de la classe $K$

4. Si no hi ha supressions secundàries als marginals de les subtaules de la classe $K$, incloure totes les supressions secundàries de les cel·les internes de les taules com a cel·les de supressió primària per a la classe $K + 1$.

5. Si hi ha supressions secundàries als marginals de com a mínim una subtaula, tornar a la classe $K - 1$. Incloure només les supressions secundàries als marginals com a supressions primàries.

6. Repetir les passes 4 i/o 5 fins que totes les subtaules s’hagin protegit en el nivell jeràrquic més baix.
Totes les subtaules es protegeixen fent servir la descomposició de Benders descrita a l’apartat 8.2.

La passa 5 no està descrita a [De 02] o [De 99]. Sense la consideració d’aquesta passa addicional es pot donar el cas que l’algorisme no sigui capaç de trobar una solució al problema.

Per tal d’explicar amb més detall com funciona l’algorisme, s’aplicarà el mètode sobre una taula de dues dimensions amb dues variables jeràrquiques. La idea s’extén facilment al cas de tenir taules amb més dimensions i un nombre arbitrari de variables jeràrquiques.

Les variables que s’usaran per il·lustrar l’exemple són dues variables categòriques relacionades amb la temàtica del projecte: una variable R (Regió) i una variable TP (Tipus de Pensió).

L’estructura jeràrquica de la variable R està il·lustrada a la figura 8.2 i la jerarquia de TP a la figura 8.1. La variable R té 4 nivells: el nivell 0 correspon al total de Catalunya, el nivell 1 correspon als àmbits del pla territorial (8 àmbits), el nivell 2 correspon a les comarques (41 comarques) i el nivell 3 als municipis (947 municipis). La variable TP té 3 nivells: el nivell 0 correspon al total de pensions, el nivell 1 està format pels tres grans tipus de pensions i el nivell 2 desglosa els grans tipus en tipologies més específiques.

![Figura 8.1. Estructura jeràrquica de la variable TP](image)

**4.1. Funcionament del procediment “top-down”**

La idea bàsica darrera el procediment “top-down”, és a dir de dalt a baix, és començar calculant les supressions secundàries per els nivells més alts de les variables jeràrquiques. En teoria la primera taula que cal protegir és la que s’aconsegueix creuant el nivell 0 de la variable R i el nivell 0 de la variable TP, és a dir el total. L’interior de la taula protegida es transporta als marginals de les taules que obtenim creuant nivells inferiors de les dues variables. A aquests marginals se’ls hi assignen pesos addicionals molt alts en els càlculs de les supressions secundàries per
Figura 8.2. Estructura jerárquica de la variable R
4. EL MÈTODE HITAS

a nivells inferiors\(^2\). Si només s’introduïxen supressions a l’interior de les subtaules no hi ha necessitat d’introduir canvis en les estructures de nivell superior.

**4.2. Construcció de les agrupacions de classes**

Per a la construcció de les agrupacions de classes, primer cal definir uns grups bàsics. Cada grup bàsic s’obté a partir de l’encreuament de diferents nivells de les variables explicatives. Les classes de taules aleshores s’obtenen a partir dels grups de tal manera que el nombre de nivells sumats és un valor constant. La taula 8.1 conté totes les classes i els grups que es poden definir a l’exemple usat.

<table>
<thead>
<tr>
<th>Classe</th>
<th>Grups</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>00</td>
</tr>
<tr>
<td>1</td>
<td>01, 10</td>
</tr>
<tr>
<td>2</td>
<td>02, 20, 11</td>
</tr>
<tr>
<td>3</td>
<td>12, 21, 30</td>
</tr>
<tr>
<td>4</td>
<td>22, 31</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
</tr>
</tbody>
</table>

**Taula 8.1. Les classes definides al creuar R amb TP**

Per exemple, el grup 31 s’obté creuant el nivell 3 de la variable R amb el nivell 1 de la variable TP. Com que el nivell 3 de la variable R correspon als municipis de Catalunya, que s’agrupen en les 41 comarques (nivell 2), el grup 31 té 41 taules on per a cada comarca es desglosa el tipus de pensions segons els tres grans grups. És a dir, el nombre de taules en un grup ve determinat pel nombre de categories “pare” que té la jerarquia en el nivell superior.

Si es defineixen les classes d’aquesta manera, els marginals de les taules en la classe \( \mathcal{K} \) s’hauran tractat com a cel·les interiors en les taules de les classes anteriors. Això significa que, cada taula a la classe \( \mathcal{K} \) es pot tractar de manera independent a les altres taules de la seva mateixa classe sempre que s’hagin tractat abans les taules de les classes de numeració inferior.

\(^2\)\text{[De 02]}\|\text{[De 99]} usen una primera versió més restrictiva del mètode on els marginals es “fixen” i per tant no es permeten supressions en els marginals.
Capítol 9
El paquet sdcTable

1. Característiques del paquet sdcTable

Bernhard Meindl, el creador i qui manté el paquet estadístic sdcTable, presenta en el seu article [Mei11a] l’estructura bàsica del paquet dins de la “Joint UNE-CE/Eurostat work session on statistical data confidentiality”, que es porta a terme a Tarragona a l’octubre de 2011. Tot i que ja s’han fet públiques versions més noves del software, la seva estructura bàsica no ha canviat.

Les primeres versions del programa estaven escrites usant programació S3. La gran diferència a partir de finals de 2011 és la reestructuració del paquet de manera que s’abandona la programació S3 i es comença a estructurar la programació segons el paradigma S4. Al programar usant S4 el que s’aconsegueix és treballar usant la programació orientada a objectes.

Les classes en S4 permeten crear objectes de dades que estan formats per un o més elements. Cada element correspon a una estructura de dades, ja sigui de tipus primitius o d’altres classes prèviament definides.

Simplifica enormement la feina de validar les dades, ja que per defecte es comprova que els elements corresponguin a les estructures de dades definides. També permet fer comprovacions de validesa addicional, de manera que es compleixin restriccions addicionals. Aquestes funcions de validació es criden cada cop que es crea un element nou. D’aquesta manera s’aconsegueix un software més robust.

Es creen funcions o mètodes que només poden ser aplicats a determinades classes. D’aquesta manera no cal fer comprovacions addicionals, ja que els elements d’entrada sempre seran correctes.

Només unes poques funcions s’exporten i són visibles als usuaris. Les funcions visibles als usuaris es mostren a la il·lustració 9.1

- La funció makeProblem() s’usa per crea objectes de la classe sdcProblem que és la base per a les següents etapes. Els valors assumits en la creació de l’objecte, com per exemple les fites inferiors i superiors, es poden modificar mitjançant la funció setInfo(). L’objecte resultant de cridar satisfactoriament...
aquesta funció és de la classe `sdcProblem`. Els inputs que s’han de passar a la funció són:

- **Dades.** Se li poden passar tant les microdades com les dades ja pre-agregades. En cas de proporcionar les dades agregades cal incloure tots els nivells de la jerarquia.
- **Informació de l’estructura de les variables que defineixen la taula.** Cal donar la informació de tots els nivells possibles i de les seves relacions de jerarquia en un format pre-definit.
- **Informació addicional.** Pesos mostrals, si és el cas.
- **Índexs de les variables que corresponen a les variables categòriques que formaran la taula.**
- **Índex de les variables de freqüències, si tenim taules de recomptes.**

- La funció `primarySuppression()` s’aplica únicament sobre elements de la classe `sdcProblem`. Especificant els paràmeters de la funció es pot escollir una regla de sensibilitat. Les actualment disponibles són la regla de la freqüència mínima, la regla del p% i la regla de la (n,k)-dominiància. Un paràmetre addicional permet indicar si el valor zero es considera sensible o no en el cas de fer servir la regla de la freqüència mínima. L’objecte resultat de la creació d’aquesta funció és també un objecte de la classe `sdcProblem`. Mitjançant la funció `setInfo()` és possible determinar manualment altres cel·les sensibles.

- La funció `protectTable()` s’aplica sobre elements de la classe `sdcProblem` que ja contenen la informació de les supressions primàries a realitzar. Aquesta funció s’usa per a solucionar el problema de la supressió secundària de cel·les. Per defecte, considera que una cel·la està protegida si no es pot calcular el seu valor de manera exacta (és a dir, suposa que els límits inferior i superior són exactament el valor de la cel·la). Els límits inferiors i superiors es poden
canviar mitjançant la funció `setInfo()`.

Després d’una execució exitosa de la funció s’obté com a resultat un element de la classe `safeObj`. La informació d’aquest objecte s’extreu mitjançant la funció `getInfo()`.

**2. Les regles disponibles per a realitzar supressions primàries**

Els tipus de regles implementades són:

- `freq`: aplica la regla de la freqüència mínima.
- `nk`: aplica la regla de la 
  \((n, k)\)-dominància.
- `p`: aplica la regla del \(p\)%.

La \((n, k)\)-dominància i el \(p\)% només es poden aplicar si a la funció `makeProblem()` s’han fet servir microdades per a generar l’element de la classe `sdcProblem`.

Els paràmetres addicionals que es poden usar en el cas d’aplicar la regla de la freqüència mínima són:

- `maxN`: paràmetre numèric que indica el nombre màxim de contribuents a una cel·la al aplicar la regla de la freqüència mínima. El seu valor per defecte és 3.
- `allowZeros`: variable lògica que especifica si les caselles buides (recompte = 0) s’han de considerar sensibles o no. El valor per defecte de `allowZeros` és “FALSE”, de manera que les cel·les buides no es consideren cel·les sensibles primàries.

**3. Estats d’anonimització de les cel·les**

Quan es crea un objecte de la classe `sdcProblem` s’assigna un estat d’anonimització a totes les cel·les. Els possibles estats que poden prendre són:

- `'u'`: la cel·la és una supressió primària i necessita ésser protegida.
- `'x'`: la cel·la s’ha suprimit de manera secundària.
- `'s'`: la cel·la pot ser publicada.
- `'z'`: la cel·la no pot ser suprimida.

Malgrat que no hi ha informació sobre el tipus ‘z’ de cel·les ni en la documentació del paquet ni en la “Vignette”, apareixen aquest tipus de cel·les quan es realitza una supressió primària amb els valor per defecte. S’ha pogut observar, però, que es correspon amb el nombre de cel·les buides.

**4. Els mètodes disponibles de resolució**

La funció `protectTable()` disposa de quatre mètodes de protecció:
OPT: correspon al mètode òptim aplicant descomposició de Benders. La documentació del paquet ja adverteix que només s’ha d’usar per a problemes de petites dimensions.

HITAS: aplica el mètode HiTas.

HYPERCUBE: aplica el mètode de l’hipercub.

SIMPLEHEURISTIC: aplica una heurística simple per a protegir les taules.

A més a més de triar el mètode de protecció de la taula, `protectTable()` disposa de dos paràmetres d’ús general:

- **verbose**: paràmetre lògic que determina si es mostra o no per pantalla informació addicional sobre la protecció de la taula. El seu valor per defecte és “FALSE”. Per a la protecció de taules grans no seria massa recomanable activar aquesta opció, ja que la quantitat de sortides per pantalla pot ser força gran.
- **save**: paràmetre lògic que determina si es guarden o no en el directori de treball els resultats temporals. El seu valor per defecte és “FALSE”.

### 4.1. El mètode de l’Hipercub implementat a sdctable

Els paràmetres específics que `protectTable()` disposa per el cas del mètode de l’hipercub són:

- **protectionLevel**: valor numèric que especifica el nivell requerit de protecció en el procediment de l’Hiper Cuba. Pren per defecte el valor 80.
- **suppMethod**: vector de caràcters de llargada 1 que defineix la regla a partir de la qual es selecciona el cub “òptim” per a una cel·la sensible. Les opcions possibles són les següents:
  - **minSupps**: minimitza el nombre total de supressions secundàries addicionals. És el valor per defecte.
  - **minSum**: minimitza la suma del nombre de contribuents a les cel·les suprimides de forma addicional.
  - **minSumLog**: minimitza el logaritme de la suma del nombre de contribuents a les cel·les suprimides de forma addicional.
- **suppAdditionalQuader**: paràmetre lògic. Si el seu valor és “TRUE” es soluciona el problema dels informants únics a una cel·la. Per defecte el seu valor és “FALSE”.

### 4.2. El mètode HiTas implementat a sdctable

Els paràmetres específics de que `protectTable()` disposa per el cas del mètode HiTas són les mateixes que les disponibles per al cas **OPT**:

- **solver**: vector de caràcters de llargada 1 que defineix el solver que es farà servir. Actualment la única opció disponible és “glpk”.
- **timeLimit**: vector numèric de llargada 1 (o NULL) que defineix el límit de temps en minuts a partir del qual l’algorisme de Benders s’hauria d’aturar i retornar una possible solució no òptima. Per defecte aquest paràmetre val NULL i és equivalent a trobar la solució òptima.
4. ELS MÈTODES DISPONIBLES DE RESOLUCIÓ

`maxVars`: vector numèric de llargada 1 (o NULL) que defineix la dimensió màxima del problema expressada en funció del nombre de variables de decisió per al qual es pot calcular una solució òptima. Si el nombre de variables que defineixen la taula és superior a `maxVars`, aleshores només es pot calcular una solució heurística no òptima. El valor per defecte d’aquest paràmetre és NULL.

`fastSolution`: variable lògica que descriu si es comença l’algorisme amb descomposició de Benders o si el que s’inicia és un procediment heurístic no òptim. No depèn del valor que tingui el paràmetre `maxVars`. Pren per defecte el valor NULL.

`approxPerc`: vector numèric de llargada 1 que defineix el percentatge respecte la fita superior proporcionada per la solució relaxada del problema mestre que s’usa com a criteri d’aturada, és a dir, que es considera suficient per acceptar que ens hem acostat prou a l’òptim. Pren per defecte el valor 10.

4.3. L’heurística simple implementada a sdcTable

A parts dels mètodes descrits al Capítol 8, que són tots mètodes ben descrits i contrastats, la funció `protectTable()` proporciona un mètode addicional al que anomena SIMPLEHEURISTIC.

Malhauradament no hi ha informació disponible sobre el mètode en la documentació del paquet sdcTable ni en cap publicació de l’autor (Bernhard Meindl).

La funció que realitza els càlculs del mètode SIMPLEHEURISTIC no es troba en el conjunt de funcions exportades als usuaris. Malgrat això, i a partir del codi font del paquet (Annex A), es pot descriure el funcionament del mètode.

Es tracta d’una heurística apropiada únicament per als casos en què la regla de detecció de cel·les sensibles sigui la de la freqüència mínima, ja que es determinen les supressions secundàries a partir de les freqüències de les cel·les.

Tenim també una premisa addicional. Tal i com passa en moltes ocasions als instituts d’estadística, l’interès és protegir els valors exactes de les cel·les sensibles. Aquesta heurística no soluciona el “problema d’auditoria” i per tant no té en compte els llindars de protecció. Tampoc soluciona el problema de les cel·les amb un únic contribuent.

A partir del codi també poder afirmar que només és un mètode apropriat en el cas que les cel·les sense contribuents no siguin considerades sensibles, ja que l’algorisme considera que aquestes caselles no es poden eliminar. De fet, el codi per a protegir una taula (Annex B) no comprova en cap moment si la opció `allowZeros` està activada o no. Per tant, en el cas que s’hagi decidit que les cel·les amb freqüència zero són sensibles, l’heurística en cap cas proporcionarà una solució al problema.

La idea bàsica rera el mètode és que per a protegir una cel·la necessitem que com a mínim hi hagi una altra supressió en la seva mateixa fila i la seva mateixa columna. Si repetim el procés per a totes les cel·les suprimides, finalment obtindrem un patró de supressió que no ens permetrà calcular els valors exactes de les caselles eliminades.
L’algorisme està dividit en tres peces: l’algorisme principal, que s’anomena `performQuickSuppression`, i dues funcions auxiliars, `suppMultDimTable` i `simpleSupp`.

### 4.3.1. La funció `performQuickSuppression`

De manera esquemàtica la funció `performQuickSuppression` treballa de la següent manera:

**Mentre** haguem introduït supressions secundàries {

Subdividim la taula en taules més petites {

**Per a** cada subaula {

Si hi ha supressions primàries o secundàries {

Tractem la taula (Cridem a la funció `suppMultDimTable`)

Canviem l’estat de les cel·les

}

Si no {

Totes les cel·les de la taula passen a tenir la categoria de ’no suprimible’

}

Si hem fet supressions {

Totes les cel·les de la taula passen a tenir la categoria de ’publicable’

Totes les cel·les de la taula de freqüència zero passen a tenir la categoria de ’no suprimible’

}

}

}

### 4.3.2. La funció `suppMultDimTable`

De manera esquemàtica la funció `suppMultDimTable` treballa de la següent manera:

Calcular totes els combinacions possibles del nombre de dimensions agafant (nombre de dimensions -1) elements sense repetició.\(^1\)

**Mentre** introduim supressions {

Per a cada element del conjunt de combinacions {

Cridar la funció `simpleSupp`

---

\(^1\)En el cas de dues dimensions tenim dues solucions: la dimensió 1 i la dimensió 2.
4.3.3. La funció **simpleSupp**

De manera esquemàtica la funció **simpleSupp** treballa de la següent manera:

Mentre introduïm supressions {
    Per a tots els elements del conjunt {
        Si hi ha més d’una casella i només una cel-la suprimida {
            Mirar les cel·les ’publicables’ i seleccionar la de freqüència més petita
            Si no hi ha cel·les ’publicables’ {
                Mirar les cel·les ’no suprimibles’ amb freqüència diferent de zero i triar
                la de freqüència menor
                Si no hi ha cel·les ’no suprimibles’ amb freqüència diferent de zero {
                    Parar. No s’ha pogut protegir la taula
                }
            }
        }
    }
}
}
Capítol 10

Resultats de l’aplicació dels mètodes de resolució del CSP

Tal i com es descriu en el Capítol 5.3 els objectius de difusió per al fitxer de pensions contributives de la Seguretat Social es divideixen en dos grups, segons el nivell de detall geogràfic al que es vulguin aplicar les tècniques de control de la revelació estadística: dades a nivell d’àmbit del pla territorial i dades a nivell municipal.

Per altra banda, abans de la resolució del problema de supressió de cel·les, cal generar un objecte de la classe sdcProblem.

1. Característiques de l’ordinador on s’han realitzat els càlculs

Tots els càlculs realitzats durant l’elaboració del projecte s’han portat a terme en un PC de les següents característiques:

- Processador Intel Core 2 Duo
- 2 GB
- Sistema Operatiu de 32 bits
- Windows Vista

La versió d’R que s’ha treballat és la 3.0.0(2013−04−03). S’ha ampliat el límit de memòria dels 1.535 Mb fins al màxim permès de 4.095 Mb. Aquest límit de memòria es degut al fet d’usar un PC amb SO Windows de 32 bits.

2. Creació del problema amb la funció makeProblem

Tal i com s’explica al Capítol 7.1, la regla usada per a determinar si una cel·la és sensible és la de la freqüència mínima amb $k = 3$.

Segons el que s’ha descrit al Capítol 9.2, si es fa servir la regla de la freqüència mínima per a determinar les cel·les sensibles, és possible realitzar els càlculs de la
generació del problema i la protecció del problema a partir tant de microdades com de dades tabulades.

S’ha estudiat el temps necessari per a la creació de les estructures de dades necessàries en amòd casos, així com el nombre de cel·les resultant.

<table>
<thead>
<tr>
<th>Variables</th>
<th># variables</th>
<th># cel·les</th>
<th>Temps segons</th>
<th>Temps min</th>
</tr>
</thead>
<tbody>
<tr>
<td>edat, prestacions, sexe, àmbit</td>
<td>4</td>
<td>5.600</td>
<td>2.006</td>
<td>33,4</td>
</tr>
<tr>
<td>import, prestacions, sexe, àmbit</td>
<td>4</td>
<td>8.680</td>
<td>2.899</td>
<td>48,3</td>
</tr>
<tr>
<td>municipi, sexe</td>
<td>2</td>
<td>3.900</td>
<td>1.410</td>
<td>23,5</td>
</tr>
<tr>
<td>municipi, prestacions, sexe</td>
<td>3</td>
<td>15.920</td>
<td>2.821</td>
<td>47,0</td>
</tr>
<tr>
<td>municipi, regim, sexe</td>
<td>3</td>
<td>11.940</td>
<td>2.274</td>
<td>37,9</td>
</tr>
<tr>
<td>municipi, prestacions, regim, sexe</td>
<td>4</td>
<td>47.760</td>
<td>6.761</td>
<td>112,7</td>
</tr>
</tbody>
</table>

TAULA 10.1. Temps de creació de l’objecte `sdcProblem` usant microdades.

<table>
<thead>
<tr>
<th>Variables</th>
<th># variables</th>
<th># cel·les</th>
<th>Temps segons</th>
<th>Temps min</th>
</tr>
</thead>
<tbody>
<tr>
<td>edat, prestacions, sexe, àmbit</td>
<td>4</td>
<td>7.200</td>
<td>34</td>
<td>0,6</td>
</tr>
<tr>
<td>import, prestacions, sexe, àmbit</td>
<td>4</td>
<td>11.160</td>
<td>47,6</td>
<td>0,8</td>
</tr>
<tr>
<td>import, edat, prestacions, sexe, àmbit</td>
<td>5</td>
<td>223.200</td>
<td>756,5</td>
<td>12,6</td>
</tr>
<tr>
<td>municipi, prestacions, regim, desagregat, sexe</td>
<td>4</td>
<td>175.120</td>
<td>589,2</td>
<td>9,8</td>
</tr>
</tbody>
</table>

TAULA 10.2. Temps de creació de l’objecte `sdcProblem` usant dades ja tabulades. El temps representat inclou el temps de càlcul de les taules en el format apropriat.

Tot i que els dos primers registres de la taula 10.2 corresponen a les mateixes entrades de la taula 10.1, el nombre de cel·les que té l’estructura del problema és diferent. Això es deu al fet que per crear l’objecte `sdcProblem` usant dades tabulades cal incloure tots els totals i subtotals derivats de les diferents estructures jeràrquiques de les variables.

Per altra banda, és força evident també la millora respecte el temps de creació quan s’usen dades ja tabulades per a generar l’objecte `sdcProblem`. És passa de tenir temps d’execució de l’ordre de minuts (o fins i tot hores), en el cas de fer servir microdades, a tenir temps d’execució de l’ordre de segons (o minuts, en els casos més grans).

Per a generar les dades d'entrada en el cas de tenir dades tabulades s’han fet servir les funcions d’R `table, addmargins, as.data.frame, rbind`. En el cas de variables amb estructura jeràrquica, cada nivell jeràrquic estava codificat en una variable diferent. El temps d’execució que apareix a la taula 10.2 inclou el temps necessari per a generar les dades d’entrada. No inclou, però, el temps de recodificació de les variables, ja que es considera que és un procés que es realitza durant el tractament previ.
3. Resultats per a dades a nivell d’àmbit territorial

Per a protegir les taules a nivell d’àmbit territorial s’han aplicat els tres mètodes disponibles al paquet sdctable que són adequats per a taules jeràrquiques i amb un nombre gran de cèl·les.

A l’hora d’avaluar els diferents mètodes s’ha pres en consideració no només el nombre de supressions secundàries que produeix cada mètode sinó també el temps d’execució necessari. Finalment, també s’ha observat el patró de supressions resultants, ja que un mètode que suprimeixi excessius totals i subtotals tampoc és adequat per a aplicar-lo a casos pràctics.

A l’Apèndix C hi ha un exemple del resum que proporciona el paquet sdctable de la solució de la protecció d’una taula. El nombre de cèl·les de la taula (# cèl·les), el nombre de supressions primàries (# ‘u’) i el nombre de supressions secundàries (# ‘x’) es poden extreure directament d’aquest resum. El temps de càlcul que proporciona el resum no es correspon amb el temps real i s’ha calculat mitjançant la funció proc.time() d’R. A la taula 10.3 s’ha emprat el mot “Simple” per referir-se a l’heurística SIMPLEHEURISTIC descrita al Capítol 9.4.

Per a les taules a nivell d’àmbit territorial s’han fet servir les següents variables:

- **àmbit**: conté 10 nivells: total, els 8 àmbits territorials i una categoria per recollir els casos en què no s’ha pogut classificar el municipi de residència del beneficiari.
- **edat**: conté un total de 20 nivells.
- **import**: conté un total de 31 nivells.
- **sexe**: conté un total de 4 nivells. Com es menciona al Capítol 7, s’ha introduït una categoria addicional per tractar els casos en els que no es té informació del sexe del beneficiari.

S’han detectat també una sèrie d’errors a l’hora de resoldre el CSP amb el paquet sdctable. Aquests errors s’han classificat en les categories següents:

- **Error1**: error de falta de memòria.
- **Error2**: error a l’hora de resoldre el MILP amb la descomposició de Benders.
- **Error3**: error d’avaluació d’una condició del codi.

Pel que fa a l’Error3, a l’Apèndix A es descriu l’error i es presenta una opció de modificació que resoldria el problema.

A la figura 10.1 estan representats els resultats de temps d’execució segons el nombre de cèl·les de la taula i el tipus de mètode resolució. La grandària dels punts correspon al nombre de supressions secundàries.
10. Resultats de l’Aplicació dels Mètodes de Resolució del CSP

<table>
<thead>
<tr>
<th>Variables</th>
<th># var.</th>
<th># cel·les</th>
<th># ’u’</th>
<th># ’x’</th>
<th>Mètode</th>
<th>Seg.</th>
<th>Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>edat, prestacions</td>
<td>2</td>
<td>140</td>
<td>6</td>
<td>8</td>
<td>Simple</td>
<td>0,5</td>
<td></td>
</tr>
<tr>
<td>edat, prestacions</td>
<td>2</td>
<td>140</td>
<td>6</td>
<td>19</td>
<td>Hipercub</td>
<td>32,9</td>
<td></td>
</tr>
<tr>
<td>edat, prestacions, sexe</td>
<td>2</td>
<td>140</td>
<td>6</td>
<td>6</td>
<td>HiTas</td>
<td>52,9</td>
<td></td>
</tr>
<tr>
<td>edat, prestacions, sexe</td>
<td>3</td>
<td>560</td>
<td>43</td>
<td>71</td>
<td>Simple</td>
<td>3,9</td>
<td></td>
</tr>
<tr>
<td>edat, prestacions, sexe</td>
<td>3</td>
<td>560</td>
<td>43</td>
<td>141</td>
<td>Hipercub</td>
<td>147,0</td>
<td>2,4</td>
</tr>
<tr>
<td>edat, prestacions, sexe</td>
<td>3</td>
<td>560</td>
<td>43</td>
<td>63</td>
<td>HiTas</td>
<td>128,6</td>
<td>2,1</td>
</tr>
<tr>
<td>import, prestacions, sexe</td>
<td>3</td>
<td>868</td>
<td>61</td>
<td>91</td>
<td>Simple</td>
<td>3,9</td>
<td></td>
</tr>
<tr>
<td>import, prestacions, sexe</td>
<td>3</td>
<td>868</td>
<td>61</td>
<td>184</td>
<td>Hipercub</td>
<td>173,7</td>
<td>2,9</td>
</tr>
<tr>
<td>import, prestacions, sexe</td>
<td>3</td>
<td>868</td>
<td>61</td>
<td>89</td>
<td>HiTas</td>
<td>169,9</td>
<td>2,8</td>
</tr>
<tr>
<td>edat, prestacions, sexe, àmbit</td>
<td>4</td>
<td>5.600</td>
<td>411</td>
<td>893</td>
<td>Simple</td>
<td>65</td>
<td>1,1</td>
</tr>
<tr>
<td>edat, prestacions, sexe, àmbit</td>
<td>4</td>
<td>5.600</td>
<td>411</td>
<td>1.761</td>
<td>Hipercub</td>
<td>3.789,6</td>
<td>63,2</td>
</tr>
<tr>
<td>edat, prestacions, sexe, àmbit</td>
<td>4</td>
<td>5.600</td>
<td>411</td>
<td>Error1</td>
<td>HiTas</td>
<td>26.476,0</td>
<td>441,3</td>
</tr>
<tr>
<td>edat, prestacions, sexe, àmbit</td>
<td>4</td>
<td>7.200</td>
<td>542</td>
<td>539</td>
<td>Simple</td>
<td>121,6</td>
<td>2,0</td>
</tr>
<tr>
<td>import, prestacions, sexe, àmbit</td>
<td>4</td>
<td>8.680</td>
<td>671</td>
<td>1.025</td>
<td>Simple</td>
<td>150,1</td>
<td>2,5</td>
</tr>
<tr>
<td>import, prestacions, sexe, àmbit</td>
<td>4</td>
<td>8.680</td>
<td>671</td>
<td>2.262</td>
<td>Hipercub</td>
<td>7.954,6</td>
<td>132,6</td>
</tr>
<tr>
<td>import, prestacions, sexe, àmbit</td>
<td>4</td>
<td>11.160</td>
<td>720</td>
<td>727</td>
<td>Simple</td>
<td>142,7</td>
<td>2,4</td>
</tr>
<tr>
<td>import, edat, prestacions, sexe, àmbit</td>
<td>5</td>
<td>223.200</td>
<td>18.210</td>
<td>8.610</td>
<td>Simple</td>
<td>23.128,3</td>
<td>385,5</td>
</tr>
</tbody>
</table>

**Taula 10.3.** Resultats del CSP per dades a nivell d’àmbit

A l’observar les taules un cop protegides es pot veure que les taules obtingudes a partir del mètode de l’hipercub tenen supressions en els totals i subtotals tot i aplicar com a criteri d’optimalitat la minimització de la suma de les freqüències de les cel·les suprimides.

4. Resultats per a dades a nivell municipal

Per a les taules de dades a nivell municipal les variables que s’han fet servir són les següents:
4. RESULTATS PER A DADES A NIVELL MUNICIPAL

Figura 10.1. Resultats del CSP per dades a nivell d’àmbit

municipi: variable municipi tractada de manera no jeràrquica. Conté 994 nivells.
municipiJ: variable municipi tractada de manera jeràrquica. El primer nivell de la jerarquia conté els 8 àmbits territorials més una categoria addicional per a recollir els municipis no classificats. El segon nivell de jerarquia correspon a les 41 comarques més un nivell addicional per als municipis no classificats. El tercer nivell de la jerarquia conté els 994 nivells de la variable municipi. En total conté 1046 nivells.
sexe: conté un total de 4 nivells.
prestacions: conté un total de 4 nivells: ’total’ i els tres grans grups de prestacions.
regim: conté 3 nivells: ’total’, ’règim general’ i ’altres règims’.
Taula 10.4. Resultats del CSP per dades a nivell de municipi

<table>
<thead>
<tr>
<th>Variables</th>
<th># var.</th>
<th># cel·les</th>
<th># 'u'</th>
<th># 'x'</th>
<th>Mètode</th>
<th>Seg.</th>
<th>Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>municipi, sexe</td>
<td>2</td>
<td>3.980</td>
<td>45</td>
<td>65</td>
<td>Simple</td>
<td>5,8</td>
<td></td>
</tr>
<tr>
<td>municipi, prestacions</td>
<td>2</td>
<td>3.776</td>
<td>166</td>
<td>132</td>
<td>Simple</td>
<td>4,7</td>
<td></td>
</tr>
<tr>
<td>municipi, prestacions</td>
<td>2</td>
<td>3.776</td>
<td>166</td>
<td>178</td>
<td>Hipercub</td>
<td>1.007,7</td>
<td>16,8</td>
</tr>
<tr>
<td>municipi, prestacions</td>
<td>2</td>
<td>3.980</td>
<td>166</td>
<td>157</td>
<td>Simple</td>
<td>6,4</td>
<td></td>
</tr>
<tr>
<td>municipi, prestacions</td>
<td>2</td>
<td>3.980</td>
<td>166</td>
<td>246</td>
<td>Hipercub</td>
<td>396,4</td>
<td>6,6</td>
</tr>
<tr>
<td>municipi, regim</td>
<td>2</td>
<td>2.985</td>
<td>29</td>
<td>33</td>
<td>Simple</td>
<td>7,1</td>
<td></td>
</tr>
<tr>
<td>municipi, regim</td>
<td>2</td>
<td>2.985</td>
<td>29</td>
<td>38</td>
<td>Hipercub</td>
<td>274,0</td>
<td>4,6</td>
</tr>
<tr>
<td>municipi, regim</td>
<td>2</td>
<td>2.985</td>
<td>29</td>
<td>26</td>
<td>HiTas</td>
<td>203,9</td>
<td>3,4</td>
</tr>
<tr>
<td>municipi, prestacions, sexe</td>
<td>3</td>
<td>15.920</td>
<td>922</td>
<td>1.242</td>
<td>Simple</td>
<td>77,2</td>
<td>1,3</td>
</tr>
<tr>
<td>municipi, prestacions, sexe</td>
<td>3</td>
<td>15.920</td>
<td>922</td>
<td>2.217</td>
<td>Hipercub</td>
<td>2.434,5</td>
<td>40,6</td>
</tr>
<tr>
<td>municipi, regim, sexe</td>
<td>3</td>
<td>11.940</td>
<td>235</td>
<td>434</td>
<td>Simple</td>
<td>44,5</td>
<td></td>
</tr>
<tr>
<td>municipi, regim, sexe</td>
<td>3</td>
<td>11.940</td>
<td>235</td>
<td>803</td>
<td>Hipercub</td>
<td>1.165,5</td>
<td>14,9</td>
</tr>
<tr>
<td>municipi, regim, sexe</td>
<td>3</td>
<td>11.940</td>
<td>235</td>
<td>Error2</td>
<td>HiTas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>municipi, prestacions, regim, sexe</td>
<td>4</td>
<td>47.760</td>
<td>3.887</td>
<td>6.292</td>
<td>Simple</td>
<td>469,5</td>
<td>7,8</td>
</tr>
<tr>
<td>municipi, prestacions, regim, sexe</td>
<td>4</td>
<td>47.760</td>
<td>3.887</td>
<td>12.763</td>
<td>Hipercub</td>
<td>14.840,4</td>
<td>247,3</td>
</tr>
<tr>
<td>municipi, prestacions, regim, sexe</td>
<td>4</td>
<td>47.760</td>
<td>3.887</td>
<td>Error2</td>
<td>HiTas</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tal com es descriu en els objectius de difusió pel que fa a les dades a nivell municipal, el que es vol es poder definir uns llindars de població per als quals poder donar una determinada informació.

A la figura 10.2 s’han representat els municipis per els quals s’ha suprimit la informació a la taula de tipus de pensió per sexe i municipi, segons el nombre d’habitants del municipi a 1 de gener de 2012. Apareix un outlier, Santa Eulàlia de Ronçana. L’aparició d’aquest outlier es deguda als problemes de codificació automàtica dels municipis. S’ha hagut de suprimir la informació d’un total de 381 municipis, tots ells de població inferior a 2.000 habitants. El total de municipis de Catalunya amb menys de 2.000 habitants és de 594.

A la figura 10.3 s’han representat els municipis per els quals s’ha hagut de suprimir la informació de la taula de règim de la Seguretat Social per sexe per municipi, segons el nombre d’habitants a 1 de gener de 2012. Segueix apareixen l’outlier de
Santa Eulàlia de Ronçana. S’ha hagut de suprimir la informació d’un total de 114 municipis, tots ells de població inferior a 2.000 habitants.

Si considerem ara la taula municipi per règim, tipus de pensió i sexe el nombre total de municipis amb més d’un 30% de cel·les suprimides és 384, tots ells de població inferior a 2.500 habitants. A la figura 10.4 s’ha representat els municipis suprimits segons el nombre d’habitants. El total de municipis amb població inferior a 2.500 habitants que hi ha a Catalunya és de 645.
10. Resultats de l'aplicació dels mètodes de resolució del CSP

Figura 10.3. Municipis no publicables de la taula Municipi per règim i sexe

Figura 10.4. Municipis no publicables de la taula Municipi per règim, tipus de pensió i sexe
Capítol 11
Conclusions

1. Objectius de difusió segura

Pel que fa als objectius de difusió s’ha pogut veure com les eines per a la resolució del CSP disponibles al paquet d’R `sdctable` poden ser una bona manera de portar a terme la protecció de les dades tabulars, tot i que la versió actualment disponible encara requeriex algunes millores.

S’ha pogut comprovar com el mètode heurístic simple proporciona solucions en temps curts per a problemes grans. Com que es compleixen les condicions necessàries per a poder aplicar el mètode pot ser una opció viable.

Pel que fa al tractament de les dades tabulars a nivell municipal s’han pogut obtenir resultats que permetin construir grups de municipis segons el seu nombre d’habitants per determinar el nivell de detall de les taules finals.

Es veu, doncs, una aplicació pràctica amb dades reals que suposen un gran nombre de registres de:

- Una aplicació dels mètodes de control de la revelació estadística a taules que podrien ser el resultat final d’un procés de producció d’estadística oficial.
- Una aplicació dels mètodes de control de la revelació estadística en les fases iniciales del procés de producció.
- Una avaluació d’un software específic per a SDC relativament nou i poc testejat fent servir dades reals.

2. Integració al sistema de producció i difusió

La finalitat última del projecte és poder integrar les metodologies testades a la producció de resultats. Actualment, R conviu amb altres eines estadístiques. Per això s’ha estudiat també la possibilitat d’integrar les anàlisis realitzats en R a altres programes d’ús habitual en els processos de producció.
2.1. Creació d’un paquet específic d’R

Com a complement del paquet específic sdcTable, i per tal d’agilitzar les anàlisis futures amb els fitxers de la mateixa naturalesa, s’han agrupat les estructures de dades necessàries i les funcions usades per exportar els resultats en un paquet d’R.

L’estructura del paquet és molt bàsica. Només pretén ser una manera senzilla d’encapsular les estructures generades que són necessàries per tal de protegir les taules que es puguin generar en diversos moments del temps.

2.2. SPSS + R

Un dels paquets estadístics d’ús habitual és l’SPSS. Malgrat que existeix paquet per a SPSS d’integració amb R, s’ha hagut de descartar aquesta possibilitat degut a una sèrie de problemes greus (veure [Cap12]):

- Només es pot usar amb la versió 2.12 d’R. Això suposa que només es poden realitzar anàlisis amb paquets d’R que estiguin compilats en aquesta versió. Per tant queden descartades les darreres versions d’sdcTable.
- No hi ha una comunicació fluida entre els dos programes.
- És lent. Les operacions que en R són pràcticament instantànies en SPSS triguen diversos segons.

2.3. SAS + R

Una altra de les eines usades de manera freqüent, sobretot pel que fa al tractament de grans conjunts de dades, és SAS. Com en el cas d’SPSS, existeix una interfície d’integració de R dins de SAS anomenada SAS/IML.

En aquest cas podem trobar a [Bew11] un llistat amb els principals inconvenients d’aquesta eina d’integració:

- Només està disponible per a ordinadors amb Windows
- En cas de tenir diverses instal·lacions d’R en l’ordinador, no hi ha opció per dir-li a SAS quina versió d’R usar.
- L’ús de grans conjunts de dades està limitat per la memòria de l’ordinador. Les dades a SAS/IML es carregen i processen a la memòria principal. El mateix passa amb R.
- R és capaç de treballar amb gran quantitat de tipus de dades, ja que tot és un objecte en R. Cada tipus de dades s’ha de convertir bé a un data.frame, bé a una matriu, per a passar-ho a SAS.

2.4. Recomanacions

Degut als inconvenients de les actuals eines d’integració el més recomanable és:

- Per a grans conjunts de dades treballar directament en R de manera remota (de la mateixa manera que es feria si es treballés amb SAS).
- Per a conjunts de dades petits i mitjans:
4. INCIDÈNCIES AMB EL SOFTWARE

- treballar directament amb R.
- usar editors de codi que tenen facilitats per a la programació en R.
- usar interferences gràfiques per a R.

3. Millora del temps d’execució

Un dels punts claus per poder aplicar a la pràctica l’ús de les tècniques de SDC és, no només la facilitat del software, sinó el temps necessari per a obtenir resultats.

En aquest sentit hi ha dos grans línies d’actuació que poden millorar molt el temps d’execució:

- Augmentar la memòria de l’ordinador amb el qual es treballa. Els 2 gigabytes del PC on es duïen a terme les executacions són fàcilment superables per als ordinadors que hi ha disponibles al mercat. Per altra banda, la darrera versió d’R permet augmentar la memòria disponible per al programa fins a quasi 4 Gb si s’usa una versió d’R de 32-bits en un sistema operatiu Windows de 64-bits o de 8Tb si s’usa un versió d’R de 64-bits.
- Millorar el codi del paquet sdcTable, ja sigui:
  - optimitzant el codi ja existent.
  - integrant codi en C++

4. Incidències amb el software

Una de les mancances del paquet és la poca informació detallada sobre el seu funcionament intern. La informació disponible a [Mei11b] i [Mei13] està més orientada a un usuari sense coneixements sobre control de la revelació estadística.

A [Mei13], hi ha un error en la descripció de la funció protectTable(). L’opció corresponent a l’heurística simple, SIMPLEHEURISTIC, es descriu amb el nom SIMPLEPROTECT.

Els errors que es produeixen durant l’execució del codi tampoc estan ben documentats. No ofereixen informació sobre l’origen de l’error i moltes vegades els missatges estan escrits en alemany.

Per altra banda, també s’ha detectat que el temps de càlcul que calcula el programa no es correspon amb el temps real de càlcul per a l’usuari. Per exemple, en la funció que permet resoldre el problema del CSP si el mètode no és l’heurística simple (calc.sdcProblem()) es realitzen diverses crides recursives sense actualitzar en cap moment el valor de la variable on s’emmagatzema el temps d’execució.

L’estructura general del paquet no compleix la majoria de recomanacions que [Gen10] fa a l’hora de crear paquets amb estructura S4.

- Hi ha funcions sobre elements de les diferents classes que no s’han declarat de manera que acceptin com a paràmetre d’entrada només els elements de la classe per a la qual han estat dissenyats.
• No s’ubiquen tots els mètodes corresponents a una classe en un únic fitxer. Cal anar navegant entre els diferents fitxers per trobar el codi de la funció que ens interessa.
• Hi ha funcions definides per a classes diferents d’elements amb una única funció. No es fa servir el polimorfisme.
• No trobem descripció de les funcions dins del codi. La majoria de descripcions dins del codi encara es troben en alemany.
• El codi no està net. Encara apareixen línies de codi comentades per a mostrar missatges per pantalla que només són útils en les primeres fases de desenvolupament del codi.
• A l’hora d’escriure ‘if statements’ no tots estan tancats amb un ‘else’. En trossos de codi molt llargs, com és el cas, això dificulta la lectura.

Pel que fa a l’eficiència del codi també és millorable. Una manera fàcil de millorar l’eficiència del codi en R és assignar l’espai necessari als objectes en lloc de definir-los i anar introduint després els seus elements.

5. Futures accions

En una segona etapa d’avaluació dels mètodes de resolució del CSP aplicats a dades reals usant `adcTable` seria interessant comprovar el funcionament de la funció `protectLinkedTables` sobre les taules realitzades als diferents nivells de desagregació territorial. D’aquesta manera es podria obtenir una bona aproximació a una protecció de totes les taules d’un pla de tabulació real feta de manera conjunta.
Apèndix A
El mètode SIMPLEHEURISTIC

1. El codi original

Codi corresponent al mètode SIMPLEHEURISTIC del paquet sdcTable.

```
1 performQuickSuppression <- function(object, input) {
2    suppMultDimTable <- function(dat, dimVars, freqInd) {
3        # protect n-dimensional table
4        simpleSupp <- function(splList, freqInd) {
5            runInd <- TRUE
6            counter <- 0
7            override <- FALSE
8            while(runInd) {
9                runInd <- FALSE
10               counter <- counter + 1
11               #cat("run:", counter,"\n")
12               for (i in 1:length(splList)) {
13                   allOk <- all(splList[[i]]$sdcStatus %in% c("z"))
14                   if (!allOk & nrow(spl[[i]]) > 1 & length(which(splList[[i]]$sdcStatus %in% c("u", "x"))) == 1) {
15                       #cat("we need to doe something: i="i,"\n")
16                       runInd <- TRUE
17                       ind.x <- which(splList[[i]]$sdcStatus == 's')
18                       f <- splList[[i]][, freqInd]
19                       toSupp <- ind.x[order(f[ind.x], decreasing=FALSE)]
20                       #cat("toSupp:", toSupp,"\n")
21                       if (is.na(toSupp)) {
22                           cat("Problem bei i="i,"\n")
23                           ind.x <- which(splList[[i]]$sdcStatus %in% c('s', 'z') & splList[[i]]$freq != 0)
24                           f <- splList[[i]][, freqInd]
25                           toSupp <- ind.x[order(f[ind.x], decreasing=FALSE)]
26                        } #endif
27                    } #endfor
28                } #endwhile
29            } #endsimpleSupp
30        } #endsuppMultDimTable
31    } #endfunction
32} #endfunction
```
override <- TRUE

if ( splList[[i]]$freq[toSupp]==0 ) {
  stop("Fehler!\n")
}

splList[[i]]$sdcStatus[toSupp] <- 'x'

s <- do.call("rbind", splList)
rownames(s) <- NULL
suppsAdded <- TRUE
if ( counter == 1 ) {
  suppsAdded <- FALSE
}
return(list(s=s, suppsAdded=suppsAdded, override=override))

nDims <- length(dimVars)
combs <- combn(nDims, nDims-1)

runInd <- TRUE
counter <- 0
override <- FALSE
patternOrig <- dat$sdcStatus
while ( runInd ) {
  counter <- counter + 1
  suppsAdded <- rep(NA, ncol(combs))
  for ( i in 1:ncol(combs) ) {
    f <- apply(dat, 1, function(x) { paste(x[combs[,i]], collapse="-") })
    spl <- split(dat, f)
    res <- simpleSupp(spl, freqInd)
    dat <- res$s
    if ( override == FALSE & res$override == TRUE ) {
      override <- TRUE
    }
    suppsAdded[i] <- res$suppsAdded
  }
  #cat("counter:", counter, "\n")
  #cat("suppsAdded:"
  if ( all(suppsAdded == FALSE) ) {
    #cat("finished! (counter="
  )

runInd <- FALSE

pattern <- dat$sdcStatus
pattern[which(dat$sdcStatus == "s")]<- "z"
return(list(pattern=pattern, ids=dat$id, override=override))

verbose <- input$verbose
pI <- get.sdcProblem(object, type="problemInstance")
strIDs <- get.problemInstance(pI, type="strID")
dat <- data.frame(id=1:length(strIDs), strID=strIDs, freq=get.problemInstance(pI, type="freq"), sdcStatus=get.problemInstance(pI, type="sdcStatus"), stringsAsFactors=F)
indices <- get.sdcProblem(object, type="partition")$indices
dimInfo <- get.sdcProblem(object, type="dimInfo")
strInfo <- get.dimInfo(dimInfo, type="strInfo")
vNames <- get.dimInfo(dimInfo, type="varName")

for (i in seq_along(vNames)) {
  dat[, vNames[i]] <- str_sub(dat$strID, strInfo[[i]][1], strInfo[[i]][2])
}
dimVars <- 1:length(vNames)
dat <- cbind(dat[, 5:ncol(dat)], dat[, 1:4])
#freqInd <- length(vNames)+3
freqInd <- match("freq", colnames(dat))

runInd <- TRUE
while (runInd) {
  override <- FALSE
  for (i in 1:length(indices)) {
    for (j in 1:length(indices[[i]])) {
      curIndices <- indices[[i]][[j]]
      subDat <- dat[curIndices,]
      nrSupps <- length(which(subDat$sdcStatus%in%c("u","x")))
      if (nrSupps > 0) {
        if (verbose) {
          cat("group:", i, "| ")
          cat("supp: ", nrSupps, " | ")
          cat("freq: ", freqInd[i], " | ")
          cat("inds: ", curIndices, " | ")
          cat("freq: ", freqInd[j], " | ")
          cat("inds: ", curIndices, " | ")
        }
      }
    }
  }
}
cat("table",j,"/",length(indices[[i]]),"| ")
cat("nrCells:",length(curIndices),"| ")
cat("nrPrimSupps:",nrSupps,"| ")
cat("override:",override,"\n")
}

res <- suppMultDimTable(subDat, dimVars, freqInd)
matchInd <- match(curIndices, res$ids)
dat$sdcStatus[curIndices] <- res$pattern[matchInd]
if (override == FALSE & res$override==TRUE)
{
  override <- TRUE
}
else {
  ind <- which(subDat$sdcStatus=="s")
dat$sdcStatus[curIndices[ind]] <- "z"
}

if (length(which(dat$freq==0 & dat$sdcStatus%in%c("u","x"))) > 0) {
  stop("fehler2!\n")
}

if (override == TRUE)
{
  # alle zellen %in% c("s","z") müssen auf "s" gesetzt werden
  dat[dat$sdcStatus %in% c("s","z"),"sdcStatus"] <- "s"
  # new
  dat$sdcStatus[dat$freq==0] <- "z"
}
else {
  if (verbose){
    cat("finished!\n")
  }
  runInd <- FALSE
}

matchID <- match(dat$strID, strIDs)
pi <- set.proBLEMInstance(pi, type="sdcStatus", input=list(index=matchID, values=dat$sdcStatus))

object <- set.sdcProblem(object, type="proBLEMInstance", list(pi))
object
2. El codi modificat

S'ha detectat un error en el codi que fa que l'algorisme s'aturi per intentar avaluar una expressió que no és possible.

```r
Error in if(splList[[i]]$freq[toSupp]==0){
  missing value where TRUE/FALSE needed
```

Aquest error correspon a la línia 29 del codi. Per a evitar aquest error cal canviar la condició a comprovar en la línia per l'expressió següent:

```r
if(is.na(ind.x)) {
```

Si a més a més volem donar més informació sobre el tipus d'error es pot canviar també la línia 30 per:

```r
stop("I can't protect the table without suppressing zero cells")
```
Apèndix B
La funció protectTable

Codi corresponent al mètode protectTable del paquet sdcTable.

```r
protectTable <- function(object, method, ...) {
  if (!method %in% c('HITAS', 'OPT', 'HYPERCUBE', 'SIMPLEHEURISTIC')) {
    stop("valid methods are 'SIMPLEHEURISTIC', 'HITAS', 'HYPERCUBE' or 'OPT'!
"")
  }
  paraList <- genParaObj(selection='control.secondary', method=method, ...)
  if (method == 'SIMPLEHEURISTIC') {
    out <- performQuickSuppression(object, input=paraList)
  } else {
    out <- calc.sdcProblem(object, type='anonWorker', input=paraList)
  }
  safeObj <- calc.sdcProblem(out, type='finalize', input=paraList)
  safeObj
```

113
Apèndix C
Exemple de resum de la solució del CSP

---
The input data have been protected using algorithm SIMPLEHEURISTIC.
---
The algorithm ran for 10 minutes and 58 seconds.
---
To protect 18210 primary sensitive cells, 8610 cells need to be additionally suppressed.
---
A total of 196380 cells may be published.
---
Duplicate cells: Only 223200 table cells are unique, the remaining 99200 cells are duplicates.

---
Structure of protected Data
---
'data.frame': 322400 obs. of 7 variables:
  $ nomambit : Factor w/ 10 levels "Alt Pirineu i Aran",..: 1 2 4 5 6 8 9 10 7 3 ...
  $ edat.grups5: Factor w/ 20 levels "De 0 a 4 anys",..: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 ...
  $ import.cat : Factor w/ 31 levels "De 1.000,01 a 1.100,00",..: 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 ...
  $ presta2 : Factor w/ 9 levels "Favor familiars",..: 1 1 1 1 1 1 1 1 1 ...
  $ sexe : Factor w/ 4 levels "Ambdós sexes",..: 2 2 2 2 2 2 2 2 2 2 ...
  $ Freq : num 0 0 0 0 0 0 0 0 0 0 ...
  $ sdcStatus : chr "z" "z" "z" "z" "z" ...

#NULL
Índex alfabètic

algorsisme de Benders, 77

classificacions estadístiques, 21
classificació d’educació, 25
CCED, 26
CINE, 25
CNED, 26
classificació d’ocupacions, 24
CCO, 24
CNO, 24
CUIO, 24
classificacions d’activitats econòmiques, 22
CCAE, 23
CIIU, 22
CNAE, 23
NACE, 23
NAICS, 23
classificacions de béns i serveis, 23
CCPA, 24
NC, 23
PRODCOM, 23
SH, 23
TARIC, 23
genomenclatures, 24
nomenclàtor, 25
NUTS, 25
confidencialitat estadística, 8
control de la revelació estadística, 49
anàlisi de resultats, 66
dades tabulars
ajust controlat, 65
arrodoniment controlat, 65
camins mínims en xarxes, 64
CSP, 63, 75
CTA, 65
descomposició de Benders per al CSP, 77
global recoding, 63
mètode de l’hipercub, 64, 80, 90
mètode HiTas, 64, 82, 90
patró de supressió, 63, 78
protecció per intervals, 64
regla \((p,q)\), 61
regra de la dominància, 61, 89
regra de la freqüència mínima, 61, 89
regra del p\%, 61, 89
soroll multiplicatiu, 64
supressió de cel·les, 63, 75
supressió primària, 63, 89
supressió secundària, 63, 75, 89
supressió primària, 75
microdades
arrodoniment, 56
dades sintètiques, 57
data swapping, 56
emmascarament amb soroll, 54
identificadors directes, 51
mètodes no pertorbatius, 53
mètodes pertorbatius, 54
MASSC, 57
microagregació, 55
mostreig, 53
PRAM, 56
quasi-identificadors, 51
rank swapping, 56
recodificació global, 53
remostrei, 56
risc de re-identificació, 52
supressió local, 53
top bottom coding, 53
variables confidencials, 52
variables no confidencials, 52
risc de revelació, 50
dades tabulars, 60
taula k-dimensional única, 60
taules de freqüències, 60
taules de magnituds, 60
taules generals, 60
taules jeràrquiques, 60
taules lligades, 61, 80
taules positives, 60

estadística oficial, 3
ESSPROS, 37
Llei d’estadística de Catalunya, 12
Llei de la Funció Estadística Pública, 10, 19
Llei del Pla estadístic, 12, 13
programa estadístic, 6
Registre de Fitxers Estadístics, 19
secret estadístic, 13, 17
normativa catalana, 19
normativa espanyola, 18

microdades, 50
microdades protegides
PUF, 51
SUF, 51

pensions
pensions contributives, 29
beneficiaris, 38
favor de familiars, 36, 41
inacapitat permanent, 30, 41
jubilació, 32, 38, 41
mort i supervivència, 34
orfandat, 35, 41
prestació temporal de viduïtat, 35
viduïtat, 34, 41
pensions no contributives, 29
protecció de dades de caràcter personal, 15, 16

Seguretat Social
Règims d’afiliació, 36
Assegurança Escolar, 36
Règim especial de mineria del carbó, 36, 41
Règim especial de treballadors
autònoms, 36, 41
Règim especial de treballadors del mar, 36, 41
Règim General, 36, 41
Bibliografía

[AAC05+] Jesús Arango Fernández, Andrés Arroyo Pérez, José Ramón Cancelo de la Torre, José Miguel Casas Sánchez, José Ignacio García Ramos i Antonio Pascual Acosta. La organización de la estadística pública en españa, 2005.


[GR02] Sarah Giessing i Dietz Repsilber. Tools and strategies to protect multiples tables with the ghquar cell suppression engine. Dins Domingo Ferrer [Dom02], pàgines 181–192.


