QÜESTIIÓ, vol. 24, 2, p. 367-389, 2000

WINSORITZACIÓ DE LA DESPESA TURÍSTICA A CATALUNYA*

MARC SÀEZ
MARIA ANTÒNIA BARCELÓ
CARME SAURINA
GERMÀ COENDERS
Departament d'Economia, Universitat de Girona*

En el marc d'actuació de l'Institut d'Estadística de Catalunya (Idescat), el sector turístic apareix com una línia prioritària, tant des de l'àmbit de la producció estadística com des de la reflexió teòrica, i dins aquest context es pot destacar la despesa turística com a variable certament estratègica, quan es tracta d'aprofundir en temes tan determinants com poden ser el de l'impacte del turisme en el creixement econòmic o el de la comptabilitat satèl·lit del sector turístic. Això justifica abastament l'interès de l'Idescat en el desenvolupament de millores metodològiques en aquesta unitat d'anàlisi. En el present article es proposa un mètode per a imputar les dades mancants i corregir els errors sistemàtics de les enquestes destinades a mesurar la despesa dels viatges turístics. La complexitat de la despesa turística fa que aquests errors siguin sobretot per omissió. El mètode proposat consisteix en winsoritzar asimètricament els residus d'un model de regressió que inclou les principals característiques del viatge realitzat com a variables predictores de la despesa diària per persona. Per a l'estimació s'empra iterativament el mètode de mínims quadrats, cosa que es pot considerar una variant de l'algorisme EM. Els errors estàndard es calculen per remostreig amb el mètode jackknife. El procediment s'il·lustra amb dades dels viatges a Catalunya realitzats per catalans, altres ciutadans de l'Estat espanyol i estrangers durant l'any 1998.

Wintsoritzation of turism expenditure in Catalonia

Paraules clau: winsorització, regressió robusta, jackknife, valors atípics, dades mancants, despesa turística

Classificació AMS (MSC 2000): 62G09, 62G35, 91B82

⋆ Aquest estudi ha estat finançat per l'Institut d'Estadística de Catalunya (Idescat), mitjançant un conveni de col·laboració signat el juliol de 1999 amb la Universitat de Girona, sota el títol *Imputació de la despesa turística a Catalunya 1997-98*.

* Departament d'Economia. Universitat de Girona. Facultat de Ciències Econòmiques i Empresarials. Campus de Montilivi. 17071 Girona. Espanya. E-mail: msaez@gnomics.udg.es

– Rebut el març de 2000.
– Acceptat el juny de 2000.

1. INTRODUCCIÓ

El fet que la renda turística se situï vora el 9 % del PIB català (Pareta i Pérez, 1998) exigeix que es disposi de mesures vàlides de la despesa turística. En aquest sentit, des de l'any 1997 la D. G. de Turisme del Dept. d'Indústria, Comerç i Turisme de la Generalitat de Catalunya realitza «l'enquesta als visitants estrangers» (EVE, Pareta i Pérez, 1998) i l'Institut d'Estadística de Catalunya (Idescat) l'enquesta als visitants procedents de la mateixa Catalunya («enquesta dels viatges dels catalans», EVC) i de la resta de l'estat l'enquesta dels viatges dels espanyols a Catalunya («Enquesta dels Viatges dels Catalans», EVEC) i de la resta de l'estat («Enquesta dels Viatges dels Espanyols a Catalunya», EVEC). A nivell de tot l'Estat Espanyol, el Instituto de Estudios Turísticos elabora anualment l'enquesta «movimiento turístico de los españoles» sobre els viatges turístics dels residents a l'Estat, sigui amb destinació interior o a l'estranger. La mateixa institució té projectat realitzar en un futur l'enquesta «estimación del gasto turístico» per incloure també les visites a l'Estat fetes per estrangers.

L'EVE s'administra de forma personal i pregunta sobre la despesa realitzada durant el viatge en el moment que el viatger es disposa a creuar la frontera i deixar Catalunya. L'EVC i l'EVEC s'administren de forma telefònica i pregunten de forma retrospectiva per la despesa dels viatges realitzats durant els darrers tres mesos. La complexitat de la despesa és considerable, donat que es divideix en capítols diversos (viatge, allotjament, compres, restauració, lleure, etc.) que es paguen en moments diferents, a vegades en monedes diferents i, fins i tot, de forma individual o col·lectiva per part de totes les persones que viatgen en grup. Així, doncs, la tasca de l'enquestat de recordar i calcular es pot considerar difícil, sobretot quan l'enquesta es fa de forma retrospectiva (Schwarz i Sudman, 1996; Sudman, Bradburn i Schwarz, 1996), encara que l'EVC i l'EVEC donin a l'enquestat l'opció de facilitar la despesa de forma global, o desagregada per conceptes o individus.

És d'esperar, per tant, que els enquestats cometin errades, sobretot per omissió. D'altra banda, el percentatge de no resposta a la pregunta sobre despesa voreja el 25 %, cosa que exigeix algun procediment d'imputació. L'objectiu del present article és proposar un mètode estadístic amb el triple objectiu de corregir les despeses reportades anormalment baixes, imputar les despeses mancants, i estimar els coeficients d'un model microeconomètric de determinació de la despesa segons les característiques del viatge. El mètode s'aplica a les dades de l'EVE, l'EVC i l'EVEC per a l'any 1998.

2. PLANTEJAMENT DE L'ANÀLISI

2.1. La informació primària

Les dades procedeixen de les enquestes EVE, EVC i EVEC, totes elles referides a viatges realitzats durant l'any 1998.

L'EVE es duu a terme de forma continuada al llarg del temps. La mostra és de tipus sistemàtica i estratificada pel calendari i les vies d'entrada i sortida de Catalunya a l'estranger. Les enquestes es fan de forma personal al cap de família en el cas de viatges en grups familiars i tenen lloc a les principals sortides cap a França per carretera, als aeroports i als ferrocarrils internacionals que surten de Catalunya per Port Bou i als autobusos turístics. Pel que fa a la despesa, es pregunta de forma agregada, distingint-se només entre l'efectuada des del país d'origen i l'efectuada a Catalunya. En aquest article es considera en tot cas la despesa total. Abans de preguntar per la despesa, es demana a l'enquestat que assenyali de tots els possibles conceptes quins corresponen a algun pagament en origen o en destinació, a fi d'ajudar a l'enquestat a recordar i incloure tots els conceptes al donar la resposta (vegeu Pareta i Pérez, 1998 per més detalls).

Les EVC i EVEC es duen a terme tres vegades l'any a mostres representatives de la població catalana i de la resta de l'Estat Espanyol major de 15 anys, obtingudes per mostreig aleatori estratificat geogràficament. Les enquestes es fan de forma telefònica assistida per ordinador al domicili habitual de l'enquestat, a qui es pregunta en primer lloc si ha fet algun viatge durant un mes determinat d'entre els quatre mesos anteriors. En el cas de l'EVEC es concreta a més que la destinació principal sigui Catalunya. La despesa es pregunta de forma separada per a cada un dels viatges reportats durant el mes de referència. La despesa la pot facilitar l'enquestat per conceptes —transport, allotjament i altres— o agregada, en forma de total per tot el grup, o com a estimació de la despesa mitjana per persona (vegeu Gomà i Bas, 1998 per més detalls).

Els qüestionaris i el mètode de recollida de dades de les enquestes EVC i EVEC són gairebé idèntics. Donat que no s'esperen comportaments radicalment diferents entre els visitants catalans i de la resta de l'estat, aquestes enquestes es varen combinar a fi de disposar de tots els viatges amb destinació a Catalunya amb qualsevol origen dins de l'Estat Espanyol.

Els viatges s'han segmentat en tres grups, amb l'objectiu d'estimar models de regressió separats per la despesa:

1) Viatges amb cost explícit de l'allotjament: viatges per motiu de lleure (excepte amb allotjament a habitatge propi, habitatge de familiars o amics o de lloguer per períodes superiors als tres mesos), i viatges per altres motius amb allotjament a hotel, càmping o residència casa de pagès. En total hi ha 239 viatges amb despesa

reportada superior a 0, realitzats per 229 enquestats per les EVC i EVEC, i 3218 enquestats i viatges per l'EVE.

2) Viatges sense cost explícit de l'allotjament: viatges per visita a familiars o amics (excepte si l'allotjament és a hotel, càmping o residència casa de pagès) i viatges per altres motius amb allotjament a habitatge propi, habitatge de familiars o amics o de lloguer per períodes superiors als tres mesos. En total hi ha 240 viatges amb despesa reportada superior a 0, realitzats per 214 enquestats per les EVC i EVEC, i 311 enquestats i viatges per l'EVE.

3) Viatges de negocis: viatges de negocis o fires, amb qualsevol tipus d'allotjament, i viatges d'estudi amb allotjament a establiments hotelers amb 441 enquestats i viatges per l'EVE. Per les EVC i EVEC els viatges de negocis i estudi són massa poc nombrosos (56 en total) per estimar-ne amb fiabilitat un model per la despesa.

2.2. Selecció de les variables

La variable depenent considerada és despesa per persona i dia. El nombre de dies s'agafa com nombre de pernoctacions més una. L'enquestat podia donar la despesa total agregada, desagregada per conceptes i/o persones, o d'ambdues maneres. En cas de conflicte es prenia el valor màxim.

Entre les variables candidates a explicatives vàrem ometre les variables socioeconòmiques o demogràfiques atès que la despesa del viatge era sovint compartida entre tota una unitat familiar i és difícil de controlar quin dels membres respon l'enquesta i fins a quin punt el seu perfil descriu bé el del conjunt de la unitat familiar. Models preliminars que incloïen aquestes variables donaven estimacions no significatives i difícils d'interpretar. Per tant, les variables escollides són aquelles relacionades amb les característiques del viatge que en principi poden afectar més el cost.

Les variables explicatives es varen agrupar en classes i codificar com a binàries; les qualitatives per evitar la presència de grups amb efectius reduïts; les numèriques per a recollir possibles efectes no lineals (per exemple, una segona persona pot reduir el cost en permetre ocupar una habitació doble, en canvi, no està tan clar l'efecte de la tercera i ulteriors persones).

Després de la recodificació, la distribució de les variables explicatives es troba a les Taules 1 i 2 i es detalla tot seguit:

1) Comunitat autònoma o país de procedència, per recollir l'efecte de la distància recorreguda i de la distribució territorial de la renda. Existiria la possibilitat, no considerada en el present treball, d'una especificació alternativa que contemplés la variable

Taula 1. *Distribució dels viatges de les enquestes EVC i EVEC.*

Viatge amb cost de l'habitatge:	Amb cost explícit	Amb cost implícit
PROCEDÈNCIA		
Catalunya	24 %	13 %
Aragó	15 %	26 %
Balears	6 %	7 %
València	13 %	10 %
Madrid	17 %	20 %
Resta comunitats nord[(a)]	16 %	11 %
Resta comunitats sud[(b)]	8 %	13 %
DESTINACIÓ		
Pirineu-prepirineu	10 %	3 %
Costa Brava	22 %	12 %
Barcelona	19 %	38 %
Costa Daurada	39 %	28 %
Maresme/Garraf/Central/Lleida	10 %	20 %
DURADA VIATGE		
1 a 7 dies	41 %	43 %
1 a 3 dies (només cap setmana)	29 %	26 %
8 o més dies	31 %	31 %
MITJÀ DE TRANSPORT		
Cotxe	67 %	62 %
Autocar	15 %	8 %
Avió	8 %	9 %
Tren i altres	10 %	21 %
ALLOTJAMENT		
Hotel	64 %	
Càmping	11 %	
Apartament de lloguer	17 %	
Altres de pagament	8 %	
PERSONES COMPARTEIXEN DESPESA		
Una	26 %	35 %
Dues	34 %	39 %
3 o més	40 %	27 %
MES DE SORTIDA		
Març-juny	26 %	22 %
Juliol-setembre	57 %	48 %
Octubre-febrer	17 %	31 %
MIDA DE MOSTRA		
Viatges	239	240
Individus	229	214

[(a)] Les altres comunitats del nord inclouen Galícia, Astúries, Cantàbria, País Basc, Navarra, La Rioja i Castella-Lleó.

[(b)] Les altres comunitats del sud inclouen Extremadura, Múrcia, Castella-La Manxa, Andalusia i Canàries.

Taula 2. *Distribució dels viatges de l'enquesta EVE.*

Tipus de viatge	Amb cost explícit	Amb cost implícit	De negocis
PROCEDÈNCIA			
Altres	11 %	2 %	15 %
França	21 %	30 %	33 %
Alemanya	31 %	28 %	11 %
Holanda	9 %	5 %	3 %
Bèlgica	4 %	18 %	3 %
Itàlia	7 %	9 %	2 %
Regne Unit	9 %	3 %	9 %
Andorra	8 %	4 %	4 %
DESTINACIÓ			
Pirineu-prepirineu/Central/Lleida	3 %	5 %	15 %
Costa Brava	39 %	27 %	4 %
Maresme	6 %	3 %	2 %
Barcelona	17 %	6 %	70 %
Costa Daurada	31 %	56 %	6 %
Garraf	4 %	4 %	3 %
DURADA VIATGE			
1 a 7 dies	54 %	31 %	91 %
8 a 13 dies	13 %	13 %	5 %
14 a 20 dies	29 %	26 %	2 %
21 o més dies	4 %	31 %	2 %
MITJÀ DE TRANSPORT			
Cotxe	55 %	90 %	61 %
Autocar	28 %	1 %	1 %
Avió	14 %	8 %	34 %
Tren i altres	3 %	1 %	3 %
ALLOTJAMENT			
Hotel	74 %		94 %
Càmping	10 %		2 %
Apartament de lloguer	16 %		1 %
Habitatge propi		100 %	3 %
Altres	0,2 %		0,2 %
PERSONES QUE COMPARTEIXEN DESPESA			
Una	4 %	4 %	65 %
Dues	47 %	53 %	25 %
Tres	15 %	17 %	4 %
Quatre (o quatre o més)	21 %	16 %	6 %
Cinc (o cinc o més)	6 %	9 %	
Sis o més	8 %		
MES DE SORTIDA			
Març-maig	20 %	34 %	36 %
Juny-setembre	66 %	46 %	29 %
Octubre-febrer	14 %	20 %	35 %
MIDA DE MOSTRA			
Individus i viatges	3218	311	441

distància a l'origen com a numèrica. La variable es podria quantificar directament o a partir d'un sistema de llindars equidistants.

2) Marca turística de destinació, per recollir l'efecte de la política de preus de cada zona. Les diferents distribucions han obligat a fer agrupacions diferents per a les diferents enquestes.
3) Durada del viatge en pernoctacions. Per als catalans i espanyols, es divideixen en estades curtes (1 a 7 pernoctacions, la majoria fora del cap de setmana) estades de cap de setmana (1 a 3 pernoctacions amb sortida un divendres o un dissabte), i estades de més d'una setmana. Per als estrangers s'observen estades més llargues i l'agrupació feta ha estat d'1 fins a 7 pernoctacions, de 8 a 13, de 14 a 20 i més de 21.
4) Tipus d'allotjament excepte pels viatges sense cost explícit de l'allotjament.
5) Mitjà de transport. La categoria «altres» inclou bàsicament ferrocarril.
6) Persones implicades a la despesa, per recollir l'efecte de possibles economies d'escala. En el cas dels estrangers, la mida de la mostra ha permès un tractament més desagregat.
7) Mes de sortida, per recollir l'efecte de la temporada. Les agrupacions s'han fet de forma diferent per als estrangers.

2.3. Mètode d'estimació i detecció de valors extremadament baixos

El procediment emprat és el de semi-winsorització asimètrica dels residus de la regressió de forma iterativa per mitjà de l'algorisme EM.

La winsorització consisteix en modificar la proporció (de la mostra que correspon a les dades més extremes abans d'estimar qualsevol paràmetre de la població. Les dades més extremes no són eliminades (que equivaldria a substituir-les per algun tipus de mitjana o mitjana condicionada) sinó que són substituïdes pel valor més allunyat entre els no extrems. S'entén així que una despesa reportada extremadament baixa correspon a un individu amb una despesa més baixa que la mitjana condicionada d'individus comparables, però no tan baixa com el valor extrem que reporta.

La winsorització formaria part, doncs, de la gran família de mètodes d'estimació de regressió resistents a valors atípics que es descriu a Chen i Dixon (1972), que pretenen reduir l'augment que es produeix en el biaix i/o la variància de les estimacions quan una part de les dades està contaminada. En aquest cas, el tipus de contaminació esperat consisteix en oblidar de reportar part de la despesa, la qual cosa ens porta a considerar la winsorització asimètrica (e.g. Kimber, 1983), tot modificant només les dades extremes de la part inferior.

Considerem la variant més operativa anomenada semi-winsorització (Guttman i Smith, 1969), en la que no es modifica una proporció prefixada d'observacions sinó totes aquelles que queden fora d'uns llindars prefixats. En aquest cas, les dades winsoritzades es fan iguals al valor d'aquest llindar.

El llindar a partir del qual una despesa es pot considerar extraordinàriament baixa depèn de les característiques del viatge i del viatger, amb la qual cosa cal considerar la winsorització a partir de la distribució condicionada de la despesa a altres variables d'interès. Per tant, semiwinsoritzem els residus d'un model de regressió que es troben per sota d'una determinat nombre negatiu z de desviacions tipus residuals $\hat{\sigma}$. Donat que els residus depenen dels paràmetres estimats que es troben distorsionats per la contaminació de les dades, cal algun procés iteratiu de re-winsorització i re-estimació.

Emprem per això una variant de l'algorisme EM[1] (Little i Rubin, 1987) que a la iteració t-èssima:

1) winsoritza els residus de cada observació i segons $e_{t+1,i} = \text{màxim}\{e_{ti}, z\hat{\sigma}_t\}$
2) suma el vector de residus winsoritzats al vector de mitjanes condicionades per construir una despesa winsoritzada $\boldsymbol{y_{t+1}} = \boldsymbol{X\hat{\hat{a}}_t} + \boldsymbol{e_{t+1}}$
3) torna a estimar els paràmetres del model de regressió per mínims quadrats ordinaris (MQO):
$$\hat{\hat{\boldsymbol{a}}}_{\boldsymbol{t+1}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y_{t+1}},\ \hat{\sigma}_{t+1} = \sqrt{\boldsymbol{e}'_{\boldsymbol{t+1}}\,\boldsymbol{e_{t+1}}/gl}$$
on gl són els graus de llibertat del model.

El procés s'atura quan $\hat{\sigma}_t = \hat{\sigma}_{t+1}$.

Aquest procés es va repetir per diferents valors de z. De fet, l'elecció de z és arbitrària. Un valor igual a 1 fou finalment seleccionat i portà a augments de la despesa mitjana no condicionada per persona i dia per a les EVC i EVEC de 287 ptes. (313 ptes. per la mediana) pels viatges amb cost explícit d'allotjament, amb un 11 % de dades winsoritzades, i de 91 ptes. (354 ptes. per la mediana) pels viatges sense cost explícit d'allotjament, amb un 5 % de dades winsoritzades. Pel que fa a l'EVE, l'augment de la despesa mitjana fou de 1060 (811 ptes. per la mediana) per als viatges amb cost explícit d'allotjament, amb un 14 % de dades winsoritzades, de 572 ptes. (1743 ptes. per la mediana) pels viatges sense cost explícit d'allotjament, amb un 9 % de dades winsoritzades, i de 4670 ptes. (2110 ptes. per la mediana) pels viatges de negocis amb un 8 % de dades winsoritzades.

[1] L'algorisme EM es sol emprar per donar valors a les dades mancants en comptes de les extremes i opera directament a partir de les mitjanes condicionades.

2.4. Càlcul dels errors estàndard

Els errors estàndard calculats pel mètode MQO no són fiables, donat que:

1) Les dades de la mateixa mostra s'empren per winsoritzar la variable depenent i per estimar el model de regressió.
2) La winsorització redueix la desviació tipus residual.
3) En el cas dels espanyols i catalans, les observacions no són independents (ens trobem amb múltiples viatges realitzats per un mateix individu). De fet la situació és assimilable a un mostreig per conglomerats en la que els viatgers són les unitats mostrals primàries.

La tècnica de «jackknife» desenvolupada per Quenouille (1956) permet el tractament de la incertesa en procediments estadístics complexos com ara la winsorització i permet calcular-ne errors estàndard robustos (Tukey, 1958).

La tècnica es basa en dividir la mostra en k grups exhaustius i mútuament excloents i en repetir l'estimació (que en el nostre cas inclou tot el procés iteratiu de winsorització i ajust MQO) k vegades, ometent de la mostra cada un dels k grups. Per cada grup j-èssim omès es calcula $\hat{\hat{\boldsymbol{a}}}_{\boldsymbol{j}} = k\hat{\hat{\boldsymbol{a}}} - (k-1)\hat{\hat{\boldsymbol{a}}}_{(\boldsymbol{j})}$, on $\hat{\hat{\boldsymbol{a}}}$ és el vector de coeficients estimat per MQO amb totes les dades i $\hat{\hat{\boldsymbol{a}}}_{(\boldsymbol{j})}$ el vector de coeficients estimat amb totes les dades excepte les del grup j-èssim. L'estimació puntual s'obté com la mitjana dels k vectors $\hat{\hat{\boldsymbol{a}}}_{\boldsymbol{j}}$ i el vector dels errors estàndard d'estimació s'obté com $1/\sqrt{k}$ vegades la desviació tipus dels components dels k vectors $\hat{\hat{\boldsymbol{a}}}_{\boldsymbol{j}}$. El quocient d'ambdós es distribueix de forma aproximadament normal per valors grans de k i permet calcular p-valors.

De fet, i seguint a Therneau i Hamilton (1997), aquest procediment és equivalent a l'estimació jackknife ponderada de Hinkley (1977). Seguint Shao i Tu (1995, plana 285), sigui $\hat{\hat{\boldsymbol{a}}}$ l'estimador MQO, i $\hat{\hat{\boldsymbol{a}}}_{(\boldsymbol{j})}$ el mateix estimador després d'ometre l'observació (o grup) j-èssim $(\boldsymbol{y_j}, \boldsymbol{x_j})$. El canvi en l'estimador MQO quan l'observació j-èssima s'omet de la mostra és igual a:

$$\hat{\hat{\boldsymbol{a}}}_{(\boldsymbol{j})} - \hat{\hat{\boldsymbol{a}}} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_j}\boldsymbol{e_j}}{\mathbf{1} - \boldsymbol{h_j}}$$

on $\boldsymbol{e_j}$ és el residual j-èssim MQO i $h_j = \boldsymbol{x_j} \cdot (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x_j}$ denota l'element j-èssim de la diagonal principal de la matriu «hat» $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (veure Shao i Tu, 1995 i també Rao i Toutenberg, 1995).

Per altra banda, sigui $\boldsymbol{D}$ la matriu de residuals «leverage», també coneguts com «dfbeta», el seu element jm-èssim d'aquesta matriu, d_{jm}, és aproximadament igual al canvi de l'estimador MQO associat a la variable explicativa m-èssima, quan l'observació j-èssima és omesa de la mostra. Cain i Lange (1984) demostren que $\boldsymbol{D}$ pot ser expressat

en un model lineal com $\boldsymbol{D} = \boldsymbol{L}(\boldsymbol{X}'\boldsymbol{X})^{-1}$, on $\boldsymbol{L}$ és una matriu amb element jk-èssim. Així, la matriu $\boldsymbol{D}'\boldsymbol{D}$no és més que l'estimador robust de la matriu de covariàncies proposat per White (1980 i 1982), és a dir:

$$\boldsymbol{S_0} = \frac{1}{n}\sum_i e_j^2 \boldsymbol{x_j} \boldsymbol{x_j'}$$

Tot i que la tècnica de «bootstrap» es considera generalment superior al jackknife per a la inferència robusta (Mooney i Duval, 1993), el jackknife és més fàcilment aplicable als dissenys mostrals complexos (Fay, 1985) i en el cas del mostreig per conglomerats cal emprar com a grups les unitats mostrals primàries (Lee et al, 1989), és a dir, els viatgers en el nostre cas.

3. RESULTATS

3.1. Determinants de la despesa turística

Les Taules 3 a 5 mostren les estimacions puntuals, errors estàndard i p-valors calculats pel mètode de jackknife. En general, els errors estàndard MQO estaven fortament esbiaixats a la baixa, tal com és d'esperar amb una variable depenent leptocúrtica, amb dades dependents i amb la desviació tipus residual reduïda pel procés de winsorització.

Pel que fa a l'EVC i a l'EVEC i pel conjunt dels paràmetres, el quocient entre l'error estàndard jackknife i el MQO es 1,29 de mitjana per als viatges amb cost explícit de l'allotjament, amb un mínim de 0,85 i un màxim de 1,68; i 1,16 de mitjana per als viatges sense cost explícit de l'allotjament, amb un mínim de 0,87 i un màxim de 1,48. Pel que fa a l'EVE, el quocient entre l'error estàndard jackknife i el MQO es 1,14 de mitjana per als viatges amb cost explícit de l'allotjament, amb un mínim de 0,63 i un màxim de 1,87; 1,23 de mitjana per als viatges sense cost explícit de l'allotjament, amb un mínim de 0,67 i un màxim de 1,84; i de 1,02 de mitjana pels viatges de negocis, amb un mínim de 0,44 i un màxim de 1,51.

Per l'EVC i l'EVEC (Taula 3), el terme constant representa la despesa mitjana per persona i dia esperada en un viatge amb origen a Catalunya, destinació a marques no incloses a la Taula 3 (Maresme, Garraf, Catalunya Central i Terres de Lleida) amb una durada inferior a una setmana majoritàriament entre setmana, realitzat en automòbil particular, passant les nits en establiments hotelers, amb una sola persona compartint la despesa i amb sortida entre els mesos de març i juny.

Els coeficients de les altres categories impliquen augment o disminució de la despesa esperada quan alguna de les característiques del viatge difereixen de les recollides en el terme constant:

Taula 3. *Estimacions pel mètode de jackknife. EVC i EVEC.*

	Viatges amb cost explícit de l'allotjament			Viatges sense cost explícit de l'allotjament		
	Coef.	error est.	*p*-valor	coef.	error est.	*p*-valor
CONSTANT	3327	1603	0.037	1168	1326	0.378
Aragó	119	1362	0.930	2720	877	0.001
Balears	2444	2747	0.373	6492	2014	0.001
València	2215	1489	0.136	3624	984	<0.001
Madrid	2071	970	0.032	5209	1096	<0.001
Resta comunitats nord	1405	1030	0.172	3934	963	<0.001
Resta comunitats sud	3209	1892	0.089	3515	1012	<0.001
Pirineu-prepirineu	3806	1367	0.005	−3332	2291	0.145
Costa Brava	2828	1184	0.016	943	969	0.330
Barcelona	7119	1870	<0.001	−590	895	0.509
Costa Daurada	2529	1190	0.033	200	983	0.838
1 a 3 dies (només cap setmana)	2091	1134	0.065	1383	853	0.105
8 o mes dies	-1304	748	0.081	−2116	642	<0.001
Autocar	−1461	1108	0.187	−1239	903	0.169
Avió	−112	2221	0.959	902	1703	0.596
Tren i altres	747	1653	0.651	102	814	0.899
Càmping	−4191	779	<0.001			
Apart lloguer <3 mesos	−3272	855	<0.001			
Altres	−937	951	0.324			
Dues	72	999	0.942	−388	781	0.619
3 o més	−684	1019	0.501	−1367	678	0.043
Juliol-setembre	2355	886	0.007	1486	811	0.066
Octubre-febrer	−1858	1354	0.170	493	718	0.492
R^2 ajustat per graus de llibertat	40 %			23 %		
Desviació tipus residual	4092			3623		

1) Tal com és d'esperar, els viatges realitzats des de fora de Catalunya presenten una major despesa, sobretot pels viatges sense cost explícit de l'allotjament, pels que el transport té un major impacte en la despesa total. En general, els augments de cost més baixos amb respecte l'origen a Catalunya es donen amb orígens a València i Aragó.

2) Pel que fa a la marca turística, només trobem diferències significatives per als viatges amb cost explícit de l'allotjament, cosa que suggereix que les diferències de preus entre marques són sobretot a l'allotjament. A les quatre marques considerades, la despesa augmenta amb respecte a la referència «altres» (Maresme, Garraf, Catalunya central i Terres de Lleida), amb un valor màxim per Barcelona.

Taula 4. *Estimacions pel mètode de jackknife. Viatges no de negocis de l'EVE.*

Tipus de viatge	Viatges amb cost explícit de l'allotjament			Viatges sense cost explícit de l'allotjament		
	Coef.	error est.	*p*-valor	coef.	error est.	*p*-valor
CONSTANT	12333	1291	<0,001	12474	2412	<0,001
França	−1360	332	<0,001	−961	779	0,109
Holanda	632	385	0,051	−1933	1459	0,093
Bèlgica	1204	600	0,022	−710	754	0,174
Itàlia	337	456	0,230	−1674	825	0,022
Regne Unit	1004	574	0,040	−322	2105	0,439
Andorra	1840	635	0,002	2888	2706	0,144
Altres	15492	1140	<0,001	3715	3215	0,125
Costa del Maresme	−1432	392	<0,001	−1126	1624	0,244
Barcelona	8791	806	<0,001	2580	1759	0,072
Costa Daurada	24	250	0,462	−133	711	0,426
Costa del Garraf	2327	806	0,002	1475	1563	0,173
Pirineu-Prepirineu/Central/Lleida	27	703	0,484	−4422	1298	<0,001
8 a 13 dies	−1548	423	<0,001	−1157	878	0,095
14 a 20 dies	−3495	277	<0,001	−1364	758	0,037
21 o més dies	−2572	1152	0,013	−117	866	0,446
Autocar	2013	452	<0,001	4422	1607	0,003
Avió	4386	786	<0,001	3609	1401	0,005
Tren i altres	5286	1698	<0,001	−1420	5393	0,396
Càmping	−1731	464	<0,001			
Apartament de lloguer	−615	417	0,070			
Altres	−14018	5386	0,004			
Dues	−4533	1413	<0,001	−2249	2069	0,139
Tres	−4967	1416	<0,001	−5449	2079	0,005
Quatre (o quatre o més)	−5832	1421	<0,001	−5505	2049	0,004
Cinc (o cinc o més)	−6267	1453	<0,001	−5541	2160	0,005
Sis o més	−4016	1536	0,004			
Març-maig	390	517	0,225	−1409	1028	0,086
Juny-setembre	2223	478	<0,001	−1580	947	0,048
R^2 ajustat	51,3 %			42,7 %		
Desviació tipus residual	7782			3896		

3) La despesa diària és màxima per viatges de cap de setmana i mínima per viatges llargs, tal com era d'esperar, tot i que algunes de les diferències tenen un nivell de significació entre el 5 i el 10 per cent.

4) El mitjà de transport no té efecte significatiu.

5) El càmping i l'apartament condueixen a una despesa significativament inferior a l'hotel.

6) El viatge en grup surt més barat per persona que l'individual, tot i que només de forma significativa quan el cost de l'habitatge és implícit.
7) La sortida durant els mesos d'estiu és la que condueix a una major despesa, tot i que només de forma significativa quan el cost de l'allotjament és explícit, cosa que fa pensar que les diferències de preus per temporades es concentren en l'allotjament.

Pels viatges no de negocis de l'EVE (Taula 4), el terme constant representa la despesa mitjana per persona i dia esperada en un viatge amb origen a Alemanya, amb destinació a la Costa Brava, amb una durada inferior a una setmana, usant com a mitjà de transport el cotxe particular, passant les nits en establiments hotelers, amb una sola persona compartint la despesa i amb sortida entre els mesos de octubre a febrer.

Els coeficients de les altres categories impliquen augment o disminució de la despesa esperada quan alguna de les característiques del viatge difereixen de les recollides en el terme constant:

1) El comportament de la despesa pel que fa al país de procedència té un comportament molt coherent en el cas dels viatges amb cost explícit observant-se el menor cost quan la procedència és França i el major cost quan la procedència és la categoria altres, categoria que aplega principalment procedències d'Estats Units i de Japó. Pel que fa als viatges sense cost explícit només apareixen significatives les procedències d'Itàlia i d'Holanda, presentant ambdues costos lleugerament inferiors al país de referència, que és Alemanya.
2) Pel que fa a la marca turística de destinació cal remarcar el menor cost significatiu que suposa el Maresme respecte de la Costa Brava, així com el cost superior observat per la marca Barcelona en el cas dels viatges amb cost explícit d'allotjament. En el cas dels viatges sense cost explícit d'allotjament, les marques significatives són Barcelona, amb un cost superior respecte de la Costa Brava, i les marques de l'interior del país amb un cost inferior.
3) La variable durada del viatge mostra que el cost és mínim per a durades entre 14 i 20 dies per ambdós tipus de viatges.
4) El cost del mitjà de transport apareix més elevat tant per l'autocar com per l'avió en ambdós tipus de viatges, fet que pot indicar que el cotxe suposa un estalvi important quan el viatge és compartit per més d'una persona.
5) El càmping i l'apartament de lloguer condueixen a una despesa significativament inferior a l'hotel en els viatges amb cost explícit d'allotjament, sent el càmping més econòmic que el lloguer, tal i com és d'esperar.
6) La variable que recull el nombre de persones contemplades en la despesa té un comportament coherent en el sentit que surt més barat viatjar de manera col·lectiva que viatjar de manera individual.

Taula 5. *Estimacions pel mètode de jackknife. Viatges de negocis de l'EVE.*

Tipus de viatge	Viatges de negocis		
	Coef.	Error est.	*p*-valor
CONSTANT	14874	1628	<0,001
Alemanya	29168	5139	<0,001
Holanda	15407	6724	0,011
Bèlgica	33844	8662	<0,001
Itàlia	8938	6914	0,099
Regne Unit	7875	2626	0,001
Andorra	20330	7340	0,003
Altres	25821	4568	<0,001
Costa Brava	−6815	3653	0,031
Costa del Maresme	−683	2406	0,388
Costa Daurada	1606	5381	0,383
Costa del Garraf	−534	5903	0,464
Pirineu-prepirineu/Catalunya central/Terres de Lleida	−1818	2382	0,223
8 a 13 dies	−16485	4182	<0,001
14 a 20 dies	−49587	4137	<0,001
21 o més dies	−25625	8592	0,001
Autocar	−580	5427	0,458
Avió	16898	3999	<0,001
Tren i altres	−174	3933	0,482
Càmping	−9561	5733	0,048
Apartament de lloguer	−15472	8494	0,035
Habitatge propi	−17501	8124	0,016
Dues	−5591	2515	0,014
Tres	9	3813	0,499
Quatre (o quatre o més)	−8363	3405	0,007
Març-maig	−1017	1988	0,305
Juny-setembre	4170	2245	0,032
R^2 ajustat	67,8 %		
Desviació tipus residual	14540		

7) La sortida durant els mesos d'estiu és la que condueix a una major despesa quan el cost de l'allotjament és explícit. En el cas dels viatges sense cost explícit la tendència semblaria la inversa encara que els resultats estan en el llindar de la significació estadística. Aquest fet fa pensar que les diferències de preus per temporades es concentren de manera especial en l'allotjament.

Els viatges de negocis de l'EVE, que presentem a la Taula 5, mereixen un comentari global abans de passar a particularitzar els resultats. Aquest tipus de viatges es presenten molt més heterogenis com es pot observar en l'enorme valor obtingut per la desviació tipus residual. Creiem que és possible que sota la denominació genèrica «negocis»

s'hi apleguin viatges amb objectius i interessos molt diversos. És per això, i per desconèixer la realitat que hi ha sota cada un dels viatges analitzats que hem d'interpretar els resultats que presentem amb precaució.

En aquest cas, el terme constant representa la despesa mitjana per persona i dia esperada en un viatge amb origen a França, amb destinació Barcelona, amb una durada inferior a una setmana, usant com a mitjà de transport el cotxe particular, passant les nits en establiments hotelers, amb una sola persona compartint la despesa i amb sortida entre els mesos de octubre a febrer.

Com en els altres casos, els coeficients de les altres categories impliquen augment o disminució de la despesa esperada quan alguna de les característiques del viatge difereix de les recollides en el terme constant:

1) El comportament de la despesa pel que fa al país de procedència té un comportament coherent respecte del país referència que és França. En tots els casos presenta un valor superior, encara que cal remarcar el fet estrany que suposa l'alta despesa observada per Alemanya i Bèlgica si la comparem amb Holanda, i l'alta despesa observada per Andorra. Aquestes dades corroboren el que hem comentat abans tot i que l'enorme error estàndard trobat les fa compatibles estadísticament.
2) Pel que fa a la marca turística, només trobem diferències significatives per la marca Costa Brava, que presenta un cost significativament inferior respecte de la marca Barcelona. Aquest resultat és raonable si es pensa en la tipologia especial que representen els viatges de negocis.
3) El comportament de la variable durada del viatge presenta la mateixa tendència que en els viatges turístics comentats anteriorment, és a dir, és mínima pels viatges entre 14 i 20 dies.
4) El mitjà de transport només surt significatiu en el cas de l'avió, presentant una despesa superior al viatge de referència que és el cotxe.
5) El càmping i l'apartament de lloguer condueixen a una despesa significativament inferior a l'hotel amb la mateixa interpretació que la feta en els viatges turístics, encara que ara la despesa menor és la que correspon a l'apartament de lloguer. Cal comentar de nou que l'opció càmping, que surt significativa en l'estudi, és una opció poc creïble pels viatges de negocis típics. Aquest fet referma la nostra hipòtesi d'heterogeneïtat en els viatges analitzats en aquest apartat.
6) La variable que regula el nombre de persones contemplades en la despesa indica clarament que viatjar amb altres persones surt més barat que viatjar de manera individual.
7) La sortida durant els mesos d'estiu és la que condueix a una major despesa, cosa que fa pensar de nou que les diferències de preus per temporades es concentren bàsicament en l'allotjament.

Taula 6. *E Despesa mitjana per marca turística. Dades de l'EVC i l'EVEC.*

	Viatgers catalans		Viatgers de la resta de l'estat	
	Dades originals	Dades imputades i winsoritzades	Dades originals	Dades imputades i winsoritzades
Resultats sense ponderar				
Pirineu-prepirineu	6371	5232	5690	5245
Costa Brava	4165	4249	7339	7377
Barcelona	6527	4695	7152	7494
Costa Daurada	5242	4658	5642	5662
Altres	2393	2181	4833	4751
Resultats elevats a la població				
Pirineu-prepirineu	5874	4403	5797	5440
Costa Brava	4481	4738	6668	6690
Barcelona	5798	3519	6648	6839
Costa Daurada	4785	4506	5927	5958
Altres	2392	2033	4509	4527

El fet que algunes variables tenen efectes diferents segons el cost de l'allotjament sigui o no explícit o bé quan es tracta de viatges de negocis en el cas dels estrangers, referma la conveniència de modelar separadament els tres tipus de viatges.

S'ha de dir, però, que el coeficient de determinació assolit, almenys pel que fa a viatges sense cost explícit de l'allotjament (en especial els de la Taula 3), podria exigir certa cautela a l'hora d'interpretar els resultats.

3.2. Imputació de dades mancants i correcció de dades atípiques

En conjunt, els signes dels coeficients són majoritàriament interpretables i es poden emprar per fer els següents tipus de depuracions sobre les dades:

1) Imputació de despeses per als que no responen o donen una despesa igual a zero: per això cal calcular la despesa per persona i dia prevista pel model per un determinat tipus de viatge. Cal partir del valor del terme constant que correspon a un viatge definit com a bàsic. A aquesta quantitat cal sumar les que es troben a continuació a la taula si el viatge compleix les condicions corresponents.
2) Depuració de les dades atípiques abans de realitzar anàlisis posteriors: per això cal Winsoritzar les despeses. A la quantitat obtinguda al punt 1) cal restar la desviació tipus residual de la darrera filera de les taules per trobar la despesa mínima creïble segons el tipus de viatge. Les despeses inferiors a aquest llindar es fan iguals al llindar.

Per exemple, pel cas dels viatges de les EVC i EVEC mostrem a la Taula 6 les despeses mitjanes per marques turístiques amb les dades abans i després de la imputació de les dades mancants per la despesa prevista i la winsorització de les dades extremes per la despesa mínima creïble.

El fet que les dades mancants corresponguessin en general a viatges amb despeses previstes més baixes (4411 de mitjana) que les dades presents (6107 de mitjana) fa que per algunes marques turístiques la despesa es redueixi amb les dades depurades, tot i la winsorització.

Cal remarcar que l'elevada manca de resposta exigeix cautela a l'hora d'interpretar els resultats. Un cas específic en aquest sentit seria el de la despesa mitjana de Barcelona en relació als viatgers catalans, en la qual es produeix una modificació molt gran en passar de les dades originals a les dades imputades. Malgrat això, el biaix de les mitjanes calculades sobre dades imputades és generalment menor al que s'obté sobre les dades originals.

4. CONCLUSIONS

Des d'un punt de vista aplicat, la no resposta i l'oblit de comptar certs conceptes són fenòmens freqüents en les preguntes sobre despesa en enquestes turístiques. A les enquestes EVC i EVEC hem trobat més del 25 % de no resposta. Els fenòmens d'oblit són més difícils de quantificar, però certament existents. Si prenem dades de l'EVEC (que exclouen per tant els viatges amb origen a Catalunya que podrien tenir un cost de desplaçament mol reduït) pels viatges amb cost explícit de l'allotjament, el 10 % dels viatges amb cotxe particular i allotjament en hotel reporten una despesa per persona i dia igual o inferior a 3500 ptes. Aquesta xifra és de 1800 pel cas de cotxe i apartament de lloguer, 1700 pel cas d'autocar i hotel i 4700 pel cas d'avió i hotel, valors tots increïblement baixos. Degut a això, les diferències entre els resultats amb les dades originals o imputades i winsoritzades són substancials, tal com mostra la Taula 6.

Des d'un punt de vista estadístic, l'article empra mètodes que encaixen dins del qualificatiu genèric de regressió robusta, com són algorisme EM, la winsorització, i el jackknife. Aquestes tècniques són conegudes de fa temps i la seva novetat resideix només en el seu ús combinat i fet a mida per resoldre el problema concret de la imputació de dades mancants i la correcció de dades atípiques en enquestes sobre la despesa turística. Actualment, el problema es resol amb procediments molt més simples. Les dades mancants a les EVC i EVEC s'imputen a partir de la despesa mitjana obtinguda per viatges amb un mateix tipus d'allotjament i una mateixa destinació (Gomà i Bas, 1998). Les dades mancants a l'EVE s'imputen pel mètode «hot deck» que consisteix a manllevar el valor de la despesa de l'individu de la mostra amb característiques més semblants (Pare-

ta i Pérez, 1998). En cap de les enquestes no s'empren mecanismes formals per corregir els valors atípics, cosa que pot causar que els valors imputats n'estiguin contaminats.

Com alternativa a l'aproximació estadística presentada en aquest article, es podrien emprar mètodes qualitatius com ara el mètode Delphi (Helmer, 1966) per obtenir pressupostos de despesa mínima creïble pels tipus de viatge més freqüents a partir dels judicis d'experts. Seria interessant aleshores comparar els resultats obtinguts amb ambdós mètodes.

REFERÈNCIES

Cain, K.C. i Lange, N.T. (1984). «Approximate case influence for the proportional hazards regression model with censored data», *Biometrics*, 40, 493-499.

Chen, E.H. i Dixon, W.J. (1972). «Estimates of parameters of a censored regression sample», *Journal of the American Statistical Association*, 6, 664-671.

Fay, R.E. (1985). «A jackknifed chi-square test for complex samples», *Journal of the American Statistical Association*, 80, 148-157.

Gomà, C. i Bas, J.M. (1998). «Els viatges dels espanyols a Catalunya i dels catalans arreu», *Nota d'Economia*, 61-62, 55-79.

Guttman, I. i Smith, D.E. (1969). «Investigation of rules for dealing with outliers in small samples of the normal distribution I: estimation of the mean», *Technometrics*, 11, 527-550.

Helmer, O. (1966). *The use of the Delphi technique. Problems of educational innovations*. Santa Monica, Ca: RAND Corporation.

Hinkley, D.V. (1977). «Jackknifing in unbalanced situations», *Technometrics*, 19,285-292.

Kimber, A.C. (1983). «Trimming in gamma samples», *Applied statistics*, 32, 7-14.

Lee, E.S., Forthofer, R.N. i Lorimor, R.J. (1989). *Analyzing complex survey data*. Newbury Park, Ca: Sage.

Little, R.J. i Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Mooney, C.Z. i Duval, R.D. (1993). *Bootstrapping. A non parametric approach to statistical inference*. Newbury Park, Ca: Sage.

Pareta, E. i Pérez, M. (1998). «El turisme estranger a Catalunya», *Nota d'Economia*, 61-62, 33-53.

Quenouille, M.H. (1956). «Notes on bias in estimation», *Biometrika*, 43, 353-360.

Rao, C.R. i Toutenburg, H. (1995). *Linear models: least squares and alternatives*. New York: Springer-Verlag.

Schwarz, N. i Sudman, S. (1996). *Answering questions : methodology for determining cognitive and communicative processes*. San Francisco, Ca: Jossey-Bass.

Shao, J. i Tu, D. (1995). *The Jackkinfe and Bootstrap*. New York: Springer-Verlag.
Sudman, S., Bradburn, N. i Schwarz, N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology*. San Francisco, Ca: Jossey-Bass.
Therneau, T.M. i Hamilton, S.A. (1997). «RhDNase as an example of recurrent event analysis», *Statistics in Medicine*, 16, 2029-2047.
Tukey, J. (1958). «Bias and confidence in not-quite large samples», *Annals of Mathematical Statistics*, 29, 614.
White, H. (1980). «A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedastictiy», *Econometrica*, 48, 817-838.
White, H. (1982). «Maximum likelihood estimation of misspecified models», *Econometrica*, 50, 1-16.

ENGLISH SUMMARY

WINSORIZATION OF TOURISM EXPENDITURE IN CATALONIA*

MARC SÀEZ
MARIA ANTÒNIA BARCELÓ
CARME SAURINA
GERMÀ COENDERS
Department of Economics. University of Girona*

The tourism sector is given high priority by the Catalan Statistical Office (Idescat), both regarding theoretical developments and production of statistical data. In this context, tourism expenditure can be regarded as a key strategic variable when it comes to assessing such important issues as the impact of tourism on economic growth and satellite accounting of the tourism sector. The Idescat is therefore committed in the develpment of methodological improvements in the field. In this article a method is suggested to impute missing data and to correct systematic errors in surveys measuring tourism expenditure. The complexity of tourism expenditure makes systematic errors to consist mainly in underreporting. The suggested method consists in asymmetrically winsorizing the residuals for daily expenditure per person in a regression model that uses the characteristics of the trip as predictors. Estimation is done iteratively with a variant of the EM algorithm. Standard errors are computed by means of the jackknife resampling method. The procedure is illustrated with data on trips to Catalonia made by Catalans, other citizens of Spain and foreigners during 1998.

Keywords: winsorization, robust regression, jackknife, outliers, missing values, tourism expenditure

AMS Classification (MSC 2000): 62G09, 62G35, 91B82

⋆ This study has been supported by the Catalan Statistical Office (Idescat).
* Department of Economics. Faculty of Economics and Business. Campus of Montilivi. 17071 Girona, Spain. E-mail: msaez@gnomics.udg.es
– Received March 2000.
– Accepted June 2000.

1. INTRODUCTION

Official statistics on tourism expenditure in Catalonia are based on personal and telephone interviews for which a large incidence of non-response and underreporting problems is expected. In this article, a statistical method is suggested to correct unusually low reported expenditures and to impute missing values.

2. ANALYSIS SETUP

2.1. Primary data

The data come from 3 surveys done in 1998. The survey of foreign visitors (SFV) done by the Catalan Tourist Office (Pareta & Pérez, 1998) is conducted in person at the main border crossings and respondents are questioned on the way back home about the trip just ended. The survey of Catalans' trips (SCT) and the survey of Spaniards' trips in Catalonia (SSTC) are done by the Catalan Statistical Office (Gomà & Bas, 1998), are conducted retrospectively by telephone and respondents are questioned about all trips done during one month. The latter two surveys are methodologically comparable and are merged in one data set. Separate regression models are estimated for:

1) trips with explicit lodgement cost (e.g. hotel, camping, apartment): 239 trips with non-missing expenditure, corresponding to 229 respondents (SCT+SSTC) and 3218 trips and respondents with non-missing expenditure (SFV).
2) trips without explicit lodgement cost (i.e. staying at friends' or relatives'): 240 trips done by 214 respondents (SCT+SSTC) and 311 trips and respondents (SFV).
3) business trips of any kind: 441 trips and respondents (SFV).

2.2. Variable selection

The considered dependent variable is daily expenditure per person. Explanatory variables include characteristics of the trip which are likely to affect cost and are dummy coded: region or country of origin, destination zone, duration, lodgement, means of transport, number of people sharing expenditure and month of departure.

2.3. Estimation method and detection of extremely low values

In order to detect outliers and correct for their effect, we use asymmetric semi-winsorization (Guttman i Smith, 1969) of the residuals of a regression model. Asymmetry

is intended to correct only for underreporting in expenditure. The method is applied iteratively by means of a variant of the EM algorithm (Little i Rubin, 1987). At each iteration t:

4) the residuals for each case are asymmetrically semi-winsorized: negative residuals larger in absolute value than a prespecified threshold of z residual standard deviations are made equal to the threshold:

$$e_{t+1,i} = \max\{e_{ti}, z\hat{\sigma}_t\}$$

5) the winsorized residual vector is added to the vector of predicted values: $\boldsymbol{y_{t+1}} = \boldsymbol{X}\hat{\boldsymbol{a}}_{\boldsymbol{t}} + \boldsymbol{e_{t+1}}$
6) The regression model is estimated by ordinary least squares (OLS)

$$\hat{\boldsymbol{a}}_{\boldsymbol{t+1}} = (\mathbf{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y_{t+1}},\ \hat{\sigma}_{t+1} = \sqrt{\boldsymbol{e}'_{\boldsymbol{t+1}}\boldsymbol{e_{t+1}}/df},$$

where *df* stands for the degrees of freedom of the model.

until $\hat{\sigma}_t = \hat{\sigma}_{t+1}$.

The method was applied for different values of z. $z = 1$ was finally selected, leading to percentages of winsorized observations between 5 % and 14 % for the different surveys and types of trip.

2.4. Computation of standard errors

OLS standard errors are incorrect because winsorization tends to reduce the residual standard deviation and the data are clustered (some respondents report on more than one trip). The jackknife (Quenouille, 1956) is the preferred method for the estimation of sampling variability of complex statistical methods for clustered samples (Fay, 1985; Lee et al, 1989), if the clustering variable (i.e. the individual) is used to define the groups which are needed to apply the jackknife. The whole iterative process of section 2.3. must be jackknifed, that is, repeated for each jackknife resampling step. Downward biases of OLS standard errors with respect to jackknife standard errors were commonly between 20 % and 30 %.

3. RESULTS

3.1. Predictors of tourism expenditure

Country or region of origin affects expenditure in the expected way: cost is higher when travelled distance is longer. The effect is higher for trips without explicit lodgement cost, as travel then increases its impact on the total expense.

The destination zone also has a significant effect, thus reflecting differential pricing policies. The effect is higher for trips with explicit lodgement cost, thus suggesting that pricing policies mostly involve lodgement. In coherence, high season trips also appear significantly more expensive only for trips with explicit lodgement cost.

The means of transport only has an effect for foreign visitors, the aeroplane being the most expensive means. Within trips with explicit lodgement costs, hotels result in higher expenditures than camping sites or apartments, as expected.

Economies of scale are revealed in the sense that longer stays and sharing expenses with others tend to significantly reduce daily expenditure per person for all types of trips.

3.2. Imputation of missing values and correction of outliers

The estimated models can be used for data cleaning purposes prior to further analyses:

3) Imputation of missing or zero expenditures by means of the model predicted values.
4) Correction of outliers (expenditures lower than the predicted value minus *z* residual standard deviations) by setting them equal to this threshold.

Even simple descriptive statistics varied markedly depending on whether the raw or cleaned data were analysed.

4. CONCLUSIONS

No response and underreporting showed to have a large incidence for all three surveys. Non-response rates neighboured 25 % and the 10th percentile of reported expenditures was usually an unbelievably low amount. As a result, the application of some robust technique to impute missing values and correct for outliers becomes critical. This article suggests using a combination of well-known robust regression techniques, which includes winsorization, the EM algorithm and jackknife standard errors.