I MODELLI INTERATTIVI PER LA RACCOLTA DEI DATI SULL'ISTRUZIONE UNIVERSITARIA VIA INTERNET

Claudia Albergamo e Giovanni Salzano - Istituto Nazionale di Statistica

Abstract

Nel presente lavoro viene descritta l'innovazione di processo che ha interessato la raccolta dei dati statistici relativi agli studenti iscritti ai corsi attivati nelle università italiane. Tale innovazione, avvenuta nell'ambito del progetto SIU (Sistema Informativo Universitario), ha consentito l'automatizzazione, con l'utilizzo di modelli elettronici, della raccolta e dell'acquisizione dei dati, che fino al 1997 si svolgevano mediante supporto cartaceo. I modelli sono stati realizzati in Visual Basic per Excel e sono stati dotati di una serie di funzioni per la verifica dei dati e per il calcolo di alcuni indicatori. I questionari elettronici e le informazioni di base per la rilevazione vengono resi accessibili attraverso una pagina *World Wide Web* (*Home Page* del SIU). Una volta concluso l'inserimento delle informazioni e portate a termine con esito positivo le verifiche di coerenza previste, il software genera un file contenente i soli dati. Il file viene quindi inviato in allegato ad un messaggio di posta elettronica.
I principali risultati sono stati il miglioramento della qualità del dato, la possibilità di aumentare la quantità di informazioni richieste e la riduzione dei tempi.

Introduzione

Dal 1937 l'Istat raccoglie annualmente dalle sedi universitarie italiane informazioni sugli studenti iscritti ai diversi corsi. L'indagine è totale e riguarda circa 80 sedi universitarie.
Fino al 1997 la rilevazione si svolgeva utilizzando il supporto cartaceo, distribuito e raccolto per via postale. I questionari erano una decina, per la maggior parte somministrati alle facoltà. Nell'ambito di ogni Ateneo, per ogni modello venivano quindi compilati tanti prospetti quante erano le facoltà.
La maggior parte dei modelli presentava in fiancata l'elenco dei corsi di laurea o di diploma della facoltà, mentre la testata variava da un modello all'altro. Le informazioni rilevate erano in massima parte rappresentate da numeri interi.
Dopo essere stati compilati in triplice copia, i questionari venivano rispediti per posta all'Istat, dove erano sottoposti ad una prima revisione e quindi alla registrazione. Le informazioni estratte dai questionari venivano quindi memorizzate su file a tracciato fisso, sui quali venivano fatti girare programmi per la verifica di alcune condizioni di coerenza interna dei dati che riguardavano essenzialmente le variazioni intervenute nelle grandezze rilevate rispetto all'anno precedente. Il processo di verifica e correzione degli errori si estendeva notevolmente nel tempo, e comportava spesso numerose interazioni con il personale delle università addetto alla compilazione dei questionari. Tanto la revisione quanto l'effettuazione dei check venivano eseguiti esclusivamente al centro.

L’impiego di tecnologie “tradizionali” di raccolta, controllo e validazione dei dati comportavano numerose limitazioni:

- l’uso della posta per la distribuzione e la raccolta dei questionari, con la possibilità di ritardi nelle consegne e di smarrimenti, comportava l'estendersi della durata della rilevazione su un arco di tempo particolarmente lungo, in genere superiore ai 10 mesi;
- i vincoli di tempo e l’esigenza di non gravare troppo sugli uffici delle università, spingevano a mantenere la quantità di informazione rilevata ad un livello inferiore rispetto a quello che le esigenze espresse dall’utenza, ed in primo luogo dagli organi di governo dell’università, avrebbero richiesto;
- al rischio di errori connesso con l’uso del supporto cartaceo per la rilevazione (trascrizioni errate o non leggibili), si aggiungeva il rischio di errate digitazioni in fase di registrazione dei questionari;
- l’effettuazione del piano di controlli al centro comportava numerose interazioni con il rispondente, che allungavano i tempi di correzione dei dati, rendendo costoso il mantenimento degli standard qualitativi richiesti;

Nel 1997 l’Istat, in risposta all’accresciuto fabbisogno informativo sul sistema universitario italiano, ha proceduto alla ristrutturazione dell’indagine, introducendo significative innovazioni nella tecnologia di rilevazione, nel quadro dei lavori per il progetto SIU (Sistema Informativo delle Università orientato alla valutazione). Nell’ambito della riorganizzazione delle indagini che ne è seguita, è stato prodotto un sistema informatizzato di raccolta e trasmissione dati che si differenzia profondamente da quelli tradizionalmente utilizzati dall’Istat nelle indagini sugli enti dell’amministrazione pubblica o del settore privato.
Il nuovo sistema è basato su questionari informatici “intelligenti” che consentono una raccolta dei dati basata sull’interscambio telematico delle informazioni con le università attraverso Internet.
La metodologia impiegata, classificabile come tecnica CASI (*Computer Assisted Self Interview*), combina diversi aspetti delle tecniche WBS (*Web Based Survey*), EMS (*E-Mail Survey*), DBM (*Disk By Mail*) ed EDI (*Electronic Data Interchange*).

1. Il rispondente "università" ed il disegno della tecnologia per la nuova indagine

Ai fini del disegno della tecnologia per la rilevazione telematica dei dati presso le università si sono dovuti affrontare tre ordini di scelte, relative rispettivamente al tipo di collegamento telematico, alle modalità di trasferimento dei dati, e al tipo di software e di interfaccia utente da utilizzare per i questionari elettronici. Nell’effettuare tali scelte si è tenuto conto delle caratteristiche dell’indagine e delle dotazioni e degli *skill* dei rispondenti, nel tentativo di selezionare, tra le possibili alternative messe a disposizione dalle tecnologie informatiche e telematiche, quelle che potevano meglio adattarsi all’indagine.
A questo fine sono stati condotti sopralluoghi presso alcune università, scelte tra quelle in cui l’organizzazione delle informazioni poteva ritenersi più vicina a quella ottimale. L'obiettivo perseguito era infatti quello di garantire che il nuovo sistema si collocasse ad un livello non inferiore rispetto allo standard già raggiunto negli ambiti locali meglio organizzati, considerati i rapidi mutamenti in corso nell’università.
Dai sopralluoghi e dagli altri riscontri diretti e indiretti effettuati si è potuto osservare che:

- gli uffici statistici e i centri di calcolo delle università utilizzano i più diffusi strumenti dell’informatica personale (System 7 o Windows, MS Office);
- tutte le Università dispongono di collegamento con Internet;
- l’uso di Netscape per l’accesso a World Wide Web e di Netscape o Eudora per la gestione della posta elettronica è pressoché generalizzato;

- gli uffici statistici sono in genere collegati ai rispettivi centri di calcolo attraverso opzioni di emulazione terminale installate sui PC utilizzati, e ricevono da questi su supporto informatico le elaborazioni o i dati grezzi da elaborare per ottenere le informazioni necessarie per l'indagine Istat;
- l'uso di Excel come foglio elettronico di elaborazione dati è diffuso e apprezzato.

La prima scelta di fondo compiuta è stata quella a favore dell'utilizzo della rete GARR (Internet), alternativa che è stata preferita a quella del trasferimento mediante collegamento "punto a punto" via modem. A tale scelta ha contribuito l'evidenza che in Italia, come nel resto del mondo, la rete Internet collega oggi la totalità delle università essendosi originariamente sviluppata come "rete delle reti" della ricerca.

Per quanto riguarda le modalità del trasferimento su Internet, si è scelto di utilizzare il protocollo HTTP per la distribuzione dei questionari, delle istruzioni e del manuale di rilevazione, mentre per i dati di ritorno dalle università si è scelto di ricorrere agli attachment della posta elettronica.

Tra le diverse alternative prese in considerazione per la realizzazione dei questionari elettronici [1] è stata scelta quella offerta dal software Excel. I motivi di tale scelta sono diversi.
- In primo luogo, l'interfaccia utente di Excel risultava particolarmente adatta alla tipologia di dati da rilevare (grandi matrici di numeri interi), ed offriva caratteristiche di standardizzazione ed ergonomia tali da garantire un'uso agevole dei questionari da parte degli utenti.
- In secondo luogo, il linguaggio di programmazione Visual Basic incorporato in Excel, forniva la possibilità di integrare in un unico file dati e programmi, consentendo di dotare il questionario dell'"intelligenza" necessaria per effettuare i controlli e gestire l'interfaccia agevolando i compiti del rispondente.
- In terzo luogo il fatto che le librerie di funzioni e l'interprete del linguaggio Visual Basic for Excel fossero incorporati nell'istallazione di Excel presso le università e non dovessero, pertanto, essere trasferite insieme ai questionari, consentiva di limitare la dimensione dei file contenenti i modelli di rilevazione, compatibilmente con le esigenze del trasferimento telematico.

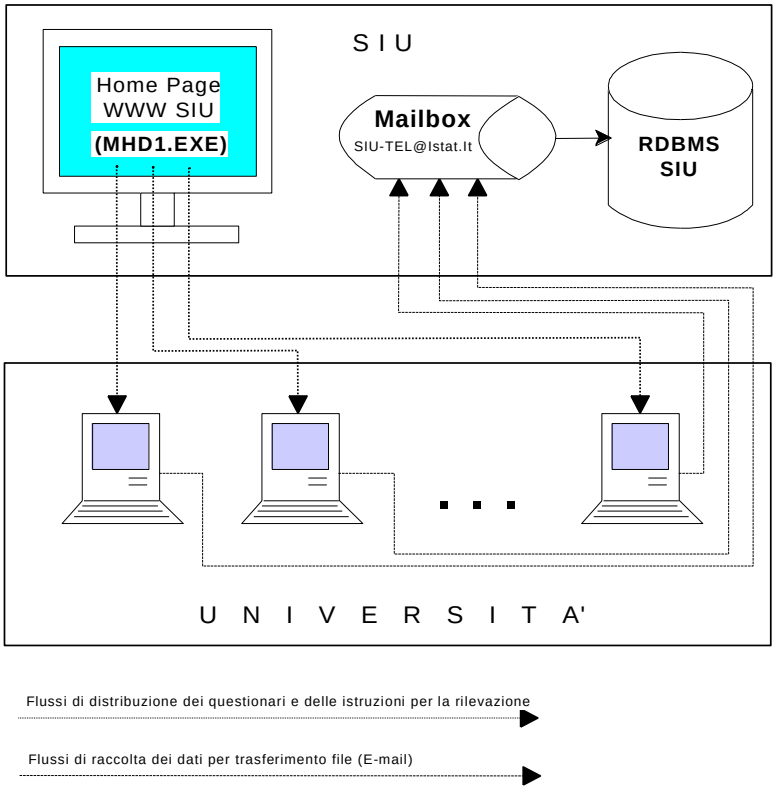
Nella scelta delle caratteristiche di cui dotare i questionari elettronici si sono anche tenute in considerazione le precedenti esperienze condotte dal Ministero dell'Università e della Ricerca Scientifica e Tecnologica (MURST), e dalla Conferenza dei Rettori delle Università Italiane (CRUI). In quest'ultima sede era stata sperimentata con un certo successo la somministrazione alle università di un questionario elettronico distribuito su dischetti e sviluppato con MS-Access.

[1] Le alternative prese in considerazione per lo sviluppo dei questionari elettronici sono state:
a) Modelli sviluppati con il database ACCESS con moduli run-time distribuibili.
b) Modelli in formato Excel contenenti programmi di selezione e controllo sviluppati con Visual Basic for Excel.
c) Uso di file ascii semplici, a tracciato fisso.
d) Modelli per Windows sviluppati con linguaggi compilati (Visual Basic, Visual C++, ecc.).
e) Form HTML per World Wide Web.
f) Programmi che consentono di gestire l'input remoto via Internet a tabelle Excel, o che consentano la compilazione remota di tabelle garantendo una interfaccia standard (applicazioni *JAVA* o *ActiveX*, e *plug-in* per *Netscape Navigator*).

2. *Il nuovo sistema di rilevazione*

Nel *nuovo sistema di rilevazione* i questionari elettronici e le informazioni di base per la rilevazione vengono resi accessibili attraverso una pagina *World Wide Web* (*Home Page* del SIU) raggiungibile attraverso la rete Internet (Cfr. fig. 1).
Le sedi universitarie vengono istruite in merito ai passi da compiere per la rilevazione, tramite una circolare distribuita sia per posta ordinaria che per posta elettronica.
I rispondenti (gli addetti degli uffici delle università incaricati della compilazione dei questionari) prelevano dalla *Home Page* del SIU il software contenente il sistema di questionari elettronici (MHD[2]) e lo fanno girare localmente, inserendo i dati ed attivando le procedure di verifica sulla base di istruzioni contenute nel software stesso.
Una volta concluso l'inserimento delle informazioni e portate a termine con esito positivo le verifiche di coerenza previste, il software genera un file contenente i soli dati, e assiste l'utente nella spedizione del file in allegato ad un messaggio di posta elettronica.
Nel caso non fosse possibile utilizzare la posta elettronica, l'utente potrà effettuare la copia del file di dati su un dischetto, che potrà essere spedito all'Istat per posta.

Fig.1. Schema dei flussi informativi del nuovo sistema di rilevazione



Oltre alle verifiche "periferiche" di errori e anomalie contestuali al momento di rilevazione dei dati effettuate dai questionari intelligenti, ulteriori e più approfonditi controlli vengono condotti centralmente appena conclusa l'importazione dei dati nell'ambito del database di produzione. Tali controlli "centrali", si avvantaggiano di maggiori risorse hardware rispetto a quelle disponibili per i controlli periferici, e di una più ampia base di dati per le verifiche di coerenza con le informazioni rilevate negli anni precedenti.

[2] Meta-acronimo per "MACCAD": Modelli Autodimensionanti a Correzione Contestuale Automatica dei Dati.

3. I nuovi modelli di rilevazione elettronici

Lo sviluppo dei nuovi *questionari elettronici*, battezzati MHD, ha perseguito tre obiettivi fondamentali:

- *minimizzare lo sforzo* che, a parità di informazioni richieste, deve essere *affrontato dal rispondente*;
- *massimizzare la qualità dei dati* rilevati;
- *minimizzare lo sforzo necessario per la manutenzione e l'aggiornamento* dei questionari.

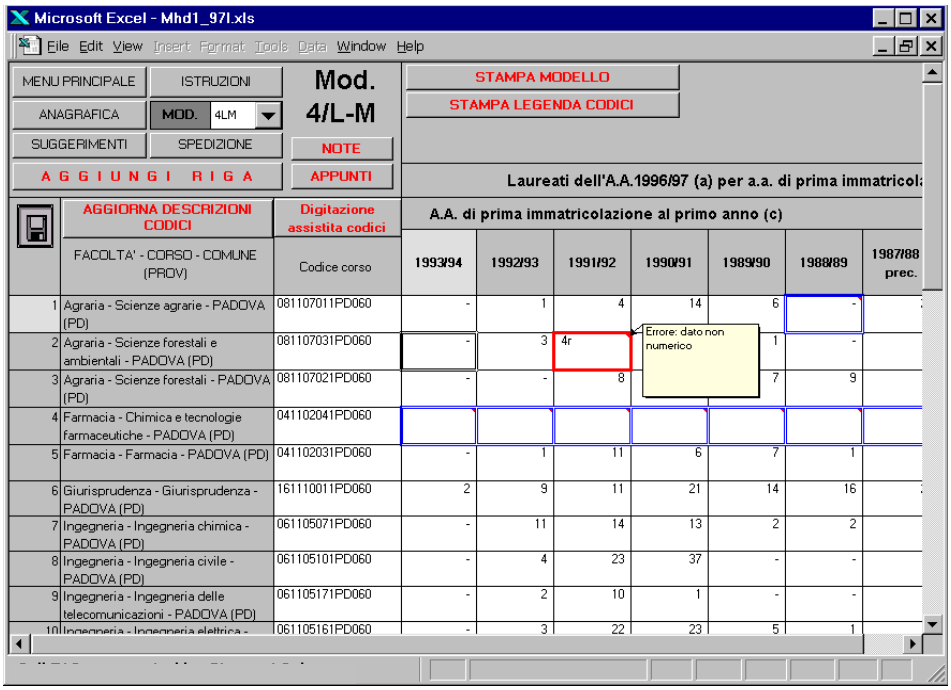
Tali obiettivi sono stati perseguiti nel rispetto di due vincoli tecnici fondamentali: quello della *dimensione minima dei file* (in coerenza con le esigenze di contenimento dei tempi di trasmissione) e quello della *compatibilità rispetto agli strumenti hardware-software* a disposizione delle università.

Per quanto riguarda l'obiettivo di *minimizzazione della molestia statistica*, si è dotato il sistema di una interfaccia utente in grado di supportare tanto l'utente medio quanto quello esperto. Un uso minimale del sistema può ridursi alla digitazione dei dati con inserimento assistito dei codici (codici ateneo, codici facoltà, codici corso di laurea), mentre un suo uso standard (e "consigliato") permette all'Ufficio statistico l'inserimento di dati ottenuti con altri sistemi (SPSS, SAS, Access, etc.) con la tecnica del "copia e incolla", consentendo di evitare la ridigitazione dato per dato delle informazioni, dove queste risultino già disponibili su supporto informatico.
I modelli vengono automaticamente precompilati con le informazioni relative ai codici ed alle denominazioni delle strutture dell'università (facoltà, corsi di laurea, corsi di diploma, etc.) risultanti dalla rilevazione precedente, evitando al rispondente di ripetere ogni anno la digitazione di tali informazioni.

Per assicurare una *migliore qualità dei dati* sono stati adottati diversi accorgimenti.
In primo luogo si sono incorporate nel questionario dettagliate istruzioni per il suo uso.
In secondo luogo sono stati implementati numerosi controlli di percorso e di coerenza dei dati immessi.
Per quanto concerne i controlli di coerenza dei dati si sono utilizzati due tipi di procedure: quelle di verifica della coerenza dei dati interni ad ogni modello e quelle di verifica della coerenza tra dati di diversi modelli.
Il primo tipo di controllo termina con l'evidenziazione delle celle del modello contenenti errori o anomalie. Passando con il mouse sopra la cella evidenziata, il programma visualizza in un riquadro una sintetica descrizione dell'errore o dell'anomalia riscontrata (Cfr. fig. 2).

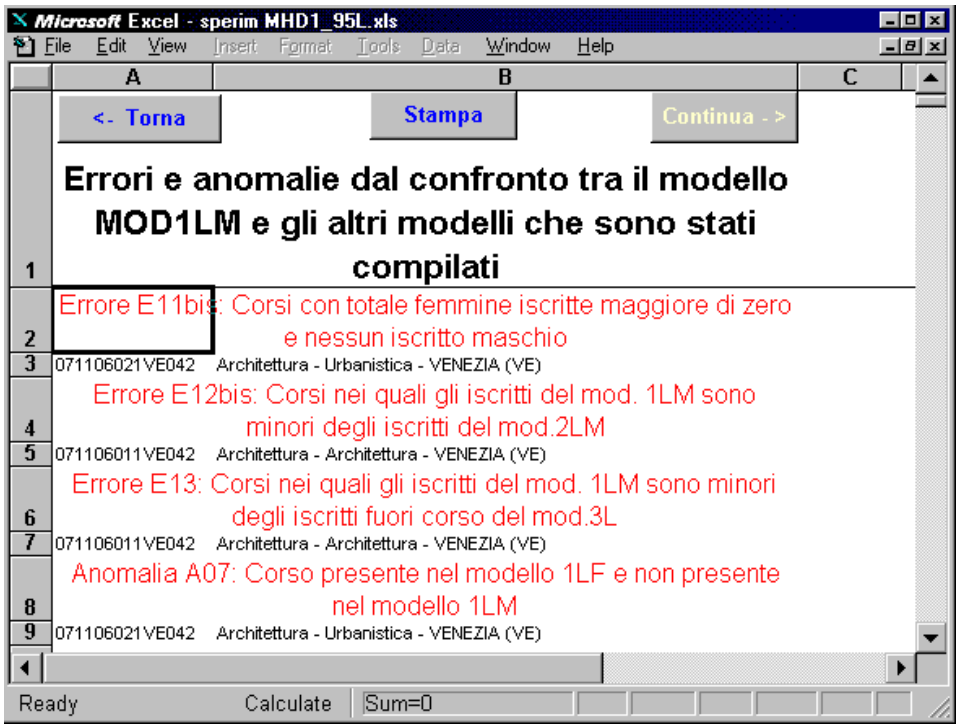
Nella figura seguente è riportata una vista del questionario relativo ai corsi di laurea, che mostra uno dei 24 modelli contenuti nel file MHD1_97L.XLS.

Fig. 2 L'intefaccia tipo di un modello elettronico MHD



Il secondo tipo di verifiche termina con la visualizzazione di un elenco degli errori e delle anomalie riscontrati, dal quale sono ricavabili i nomi dei modelli ed i codici identificativi dei record che hanno dato luogo alle anomalie e agli errori riscontrati.

Fig. 3. Elenco tipo con indicazioni di errori / anomalie tra modelli



Mentre in presenza di errori non viene permesso all'utente di completare la procedura per la spedizione dei dati inseriti, nel caso delle anomalie, pur avvertendo l'utente del fatto che la riscontrata anomalia

potrebbe essere causata da un'errata digitazione o da un errato calcolo, viene consentito di condurre a termine la spedizione.
Un ulteriore contributo alla qualità dei dati si è potuto ottenere fornendo al rispondente, alla fine del processo di rilevazione, una batteria di indicatori calcolati a partire dai dati inseriti. Questi indicatori, oltre a costituire un "premio di produzione" per il rispondente, consentono l'individuazione di eventuali situazioni anomale derivanti da errori verificatisi nell'inserimento dei dati. Gli indicatori costituiscono peraltro un risultato fondamentale del SIU in quanto Sistema Informativo Orientato alla Valutazione.

Riguardo al terzo obiettivo richiamato, si è perseguita la *massima automazione*, perseguendo l'obiettivo ideale di una indagine che si gestisca totalmente da sola. In quest'ottica i modelli sono stati resi il più possibile indipendenti dal programma, generalizzando dove possibile le routine in codice in modo da consentire che l'aggiunta o la modifica dei questionari non comporti interventi sul codice. Ogni modello è quindi stato dotato di una sorta di DNA, contenente le informazioni sul modello necessarie per la parametrizzazione delle procedure che su di esso agiscono.

Per quanto concerne il requisito della *dimensione minima*, il sistema utilizza un *principio di auto-dimensionamento* in base al quale i modelli vengono dimensionati e predisposti nelle loro parti variabili, solo dopo che il questionario è stato ricevuto dall'Università. Il file che viene trasferito dall'Istat alle Università contiene quindi modelli composti da un'unico record di meta-dati, nel quale sono riportate le informazioni di formato e le formule per i controlli.
Questa caratteristica, unita alla generalizzata diffusione di Excel [3], ha consentito di limitare notevolmente l'occupazione dei file in memoria.

La dimensione dei file MHD1_xxL.XLS ed MHD1_xxD.XLS utilizzati attualmente per l'indagine è di approssimativamente 2,5 Mbyte, mentre quella dei file in formato compresso utilizzati per la distribuzione dei questionari su Internet è di circa 0,7 Mbyte. Il file di ritorno trasferito dalle Università all'Istat contiene solo dati (non contiene cioè i programmi, le specifiche di formato ed i metadati contenuti nel file originariamente prelevato dagli uffici statistici), e si mantiene in generale al di sotto dei 100 Kbyte.

Al fine di assicurare una più ampia *compatibilità*, sono state predisposte due versioni dei questionari. Una compatibile con Excel 5.0 ed Excel 7.0, ed un'altra con Excel 97.

4. Gli esiti della prima applicazione del nuovo sistema di rilevazione

Dal maggio 1998 al modello cartaceo, usato per la rilevazione dei dati sul sistema universitario italiano, è stato affiancato il modello elettronico. Nonostante i risultati delle sperimentazioni e i sopralluoghi presso le università condotti tra il 1996 ed il 1998 conducessero a guardare con ottimismo alle possibilità di generalizzazione dell'utilizzo delle nuove tecnologie nella maggior parte delle università, si è preferito mantenere per il rispondente la possibilità di utilizzare i questionari cartacei. Le università hanno quindi potuto scegliere il modello cui ricorrere per rispondere alle richieste dell'Istituto.
L'informatizzazione della rilevazione ha riscosso un certo successo tra gli atenei: circa i due terzi delle università hanno utilizzato il nuovo sistema di rilevazione basato sui questionari elettronici MHD. Questa quota varia a seconda della collocazione geografica degli atenei. Dalla tabella 1 possiamo osservare come la quota delle università che hanno scelto di ricorrere al modello elettronico diminuiscano

[3] Nell'assunzione che la disponibilità di Excel sia generalizzata presso gli uffici statistici, sarà possibile evitare di spedire insieme al file .XLS anche moduli *run-time* del programma, come invece sarebbe necessario fare ove si fosse utilizzato ACCESS.

progressivamente passando dal Nord al Sud del paese. Nel Nord Italia ben l'83,3% dei rispondenti ha utilizzato il modello elettronico, al Centro sono stati poco più della metà (52,2%), mentre al Sud questa percentuale si riduce al 38,7%.
L'attività di assistenza tecnica fornita alle università nel corso della rilevazione, ha portato a constatare una certa disparità nel livello di informatizzazione tra le università del Sud e quelle del Nord. Questa osservazione sembrerebbe spiegare, almeno in parte, la relazione tra la collocazione geografica e il tasso di risposta, evidenziata dalla tabella 1.

Tabella 1 – Modalità di risposta degli atenei secondo la ripartizione geografica (composizione percentuale).

Ripartizioni Geografiche	Università che hanno utilizzato:	
	Modello Cartaceo	Modello Elettronico
Nord	16,7	83,3
Centro	47,8	52,2
Sud	61,3	38,7
Totale	**34,2**	**65,8**

La prima applicazione del nuovo sistema di rilevazione ha consentito di valutare il carico di lavoro aggiuntivo, rispetto al sistema tradizionale di rilevazione, dovuto all'assistenza tecnica *on-line*. Questo lavoro si è concentrato principalmente nelle prime settimane della rilevazione e nei giorni a ridosso delle scadenze della consegna dei modelli, periodi nei quali le telefonate da parte delle università hanno raggiunto un'intensità tale da richiedere l'impegno di una persona per circa metà giornata lavorativa. Gli utenti hanno richiesto assistenza principalmente per il prelievo dei programmi dalla rete e la generazione dei modelli - nella prima fase - e per il salvataggio e la spedizione dei dati - nella fase finale.
Nel corso della rilevazione la maggior parte delle richieste di intervento da parte delle università si sono concentrate su aspetti contenutistici del dato più che su aspetti tecnici, apportando, quindi, un carico di lavoro aggiuntivo minimo.
Infine l'invio dei dati via e-mail, già sottoposti ad una serie di controlli di qualità, ha permesso di eliminare la fase di registrazione e di ridurre notevolmente i controlli sui dati, riducendo sia i tempi che la quantità di lavoro necessaria per l'acquisizione dei dati.

5. Vantaggi, svantaggi e possibili miglioramenti del nuovo sistema di rilevazione

Nonostante i vantaggi del nuovo sistema di rilevazione siano molteplici ne daremo qui solo un sintetico elenco, essendo maggiormente interessati ad indagarne i limiti e le difficoltà, in vista degli ulteriori perfezionamenti che potranno essere ottenuti nelle future applicazioni.

Tra i **vantaggi** del nuovo sistema di rilevazione vi sono: la velocizzazione dei processi di acquisizione, la correzione e verifica dei dati contestualmente alla loro rilevazione, una riduzione dei costi connessi all'acquisizione delle informazioni (tanto di quelli sopportati dall'Istat quanto di quelli sopportati dal rispondente), il miglioramento della qualità dei dati, un ampliamento della quantità di informazione raccolta.

Le **difficoltà** ed i **limiti** sono riconducibili fondamentalmente a 4 categorie: difficoltà di organizzazione interna, limiti di automazione, limiti di carico telematico, limiti di complessità d'uso.

Le difficoltà **organizzative interne** scaturiscono dalla molteplicità delle competenze necessarie a gestire i diversi momenti del ciclo della nuova indagine telematica (rilevazione, revisione, check, correzione degli errori, validazione, integrazione dei dati nel sistema informativo), ed in particolare dalla necessità di affiancare le competenze statistiche a quelle informatiche (sistemistiche, di analisi e di programmazione).

Nonostante la metodologia sviluppata consenta un enorme salto in avanti nel senso di una **maggiore automazione**, continuano a sussistere margini per una possibile ulteriore informatizzazione delle procedure. In particolare, il processo di automazione potrebbe essere allargato a comprendere anche molte delle fasi che sopravvivono manuali, tra quella della raccolta, quella dell'elaborazione e quella della diffusione. Per esempio una migliore automazione potrebbe ottenersi con lo sviluppo di una procedura di "mail processing" che consenta una gestione automatica degli *attachment* contenenti i file di ritorno.

Un importante limite delle nuove procedure di rilevazione è costituito dal **grado di alfabetizzazione informatica dei rispondenti** che esse richiedono e dalla dotazione hardware e software necessaria per l'accesso e l'uso dei questionari intelligenti. La nuova metodologia richiede infatti nuovi skill per gli addetti delle università che vengono incaricati della rilevazione, oltre alla disponibilità di personal computer e di accesso alla rete Internet.