SKY: un nuovo approccio alla WEB ANALYTICS

di Silvio Gerli, **Smartstat, e** Marco Fattore,
Università degli Studi di Milano Bicocca

L'operatore televisivo ha commissionato e implementato una soluzione in grado di sfruttare le potenzialità insite nei dati raccolti dalle vendite via Internet, che sono in continuo aumento.
Un vero e proprio tesoro di informazioni, che possono generare nuove opportunità di marketing.

Con questo articolo continuiamo il percorso di approfondimento sul lavoro del data scientist attraverso il racconto di casi reali ed emblematici. Nello scorso numero abbiamo illustrato una innovativa risposta a una domanda precisa e ben posta sul controllo sintetico delle performance della forza vendita. Presentiamo qui un altro caso molto interessante, realizzato con Sky Italia, e volutamente molto differente dal precedente, per mostrare come la data science non sia una disclina limitata e confinabile in definizioni e schemi chiusi, ma anzi tenda ad occupare spazi nuovi, integrando componenti consulenziali, metodologiche e tecnologiche. Il risultato del progetto che stiamo per descrivere è stato lo sviluppo di un processo di sintesi e visualizzazione dei dati di navigazione e accesso al sito web aziendale, che supera i classici grafici a cui siamo abituati da anni, fornendo una vista dei dati organica, dinamica, più ricca di significato e più vicina alle esigenze di business.

La sfida

Già dalla richiesta iniziale del Committente, il progetto si inquadra subito come "open", richiedendo un approccio "laterale" per valorizzare la capacità innovativa della data science. Sintetizzando, con il supporto di Aldo Agostinelli, Digital Sales & Mktg Director di Sky, e di Gian Piero Zanette, Digital sales analysis manager, insieme ad altri due componenti del proprio team (Erica Verde e Fiorenza Furlan) ha ingaggiato Smartstat in partnership con Spindox con la seguente richiesta: "Il digitale per Sky rappresenta un touchpoint fondamentale e sempre più strategico. Sky permette ai propri prospect di raccogliere informazioni e sottoscrivere nuovi abbonamenti e ai propri clienti di informarsi e gestire il proprio abbonamento attraverso le property digitali fruibili da tutti i device. In vista di azioni di marketing e commerciali, quali informazioni è possibile generare a partire dalla mole di dati provenienti dalla navigazione web anonima dei nostri utenti e sistematicamente raccolti sulla piattaforma Adobe?". In realtà questa domanda non immediatamente concreta, volutamente "vaga", è espressione di un approccio lungimirante: cominciare a mettere a frutto i flussi di dati in modo sistematico, cercando di generare nuove prospettive operative. Paradossalmente, a una domanda posta in modo aperto si risponde con un forte impianto metodologico, che consenta l'esplorazione di percorsi alternativi, senza venire meno a robustezza ed efficacia. Il progetto è stato quindi impostato con i seguenti passi operativi:

1. design e strutturazione della base dati sulla quale effettuare le elaborazioni e controllo della consistenza e dell'utilizzabilità dei dati ai fini del progetto;
2. identificazione degli obbiettivi possibili con i dati analizzati;
3. identificazione e condivisione delle feature e conseguente processo di estrazione e generazione delle stesse;
4. prototipizzazione dei processi di elaborazione e di visualizzazione con alcuni esempi.

1. Design e strutturazione della base dati - Controlli di qualità

Data la natura pilota del progetto, non è stato possibile accedere direttamente ai raw-data, ma esclusivamente ai dati disponibili dalla piattaforma di Adobe, sotto forma parzialmente aggregata (per esempio, il time stamp di ciascun accesso o evento sul sito è sostituito dall'indicazione dell'intervallo temporale, di ampiezza pari a 1 ora, in cui esso è avvenuto). La base dati disponibile per le attività è quindi composta da tabelle relative a:

I. Eventi di accesso al sito Sky.
II. Eventi di navigazione successiva al primo accesso al sito.
III. Eventi di inserimento nel carrello dei pacchetti Sky.
IV. Eventi di acquisto dei pacchetti Sky.

Le verifiche di consistenza e di qualità hanno confermato la pertinenza e l'affidabilità dei dati, consentendo di passare alle fasi operative del progetto.

2. Identificazione degli obiettivi operativi

Sulla base dei bisogni espressi dal Committente, sono state identificate tre linee di progetto distinte:

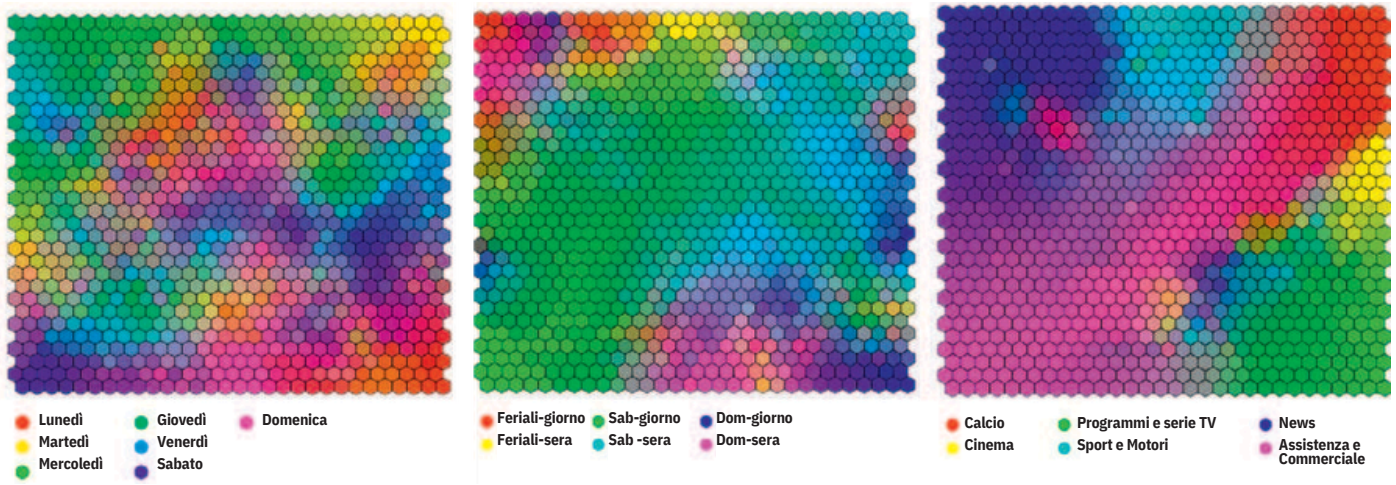
A) Fornire uno strumento sintetico descrittivo per la visualizzazione dei comportamenti di navigazione degli utenti, sul sito di Sky.
B) Individuare criteri e strumenti per l'analisi dei visitatori che popolano il proprio carrello e su coloro che acquistano pacchetti, con l'obiettivo di identificare e migliorare le azioni di autopromozione.
C) Fornire spunti per il miglioramento delle azioni di upselling.

3. Definizione ed estrazione delle feature

Normalmente, la descrizione dei comportamenti dei cookie non avviene direttamente a partire dai dati elementari, ma da un layer intermedio di informazioni desunte da quelle elementari e che esprimono tratti interessanti dei visitatori. Per esempio, dal dato elementare sulla tecnologia utilizzata per l'accesso (sistema operativo con la versione, browser con la versione, plugin installati ecc.) si possono desumere informazioni sul rapporto che il soggetto ha con la tecnologia e quindi caratterizzarne il profilo. Queste informazioni "intermedie" sono dette "feature" e costituiscono normalmente il vero layer da cui generare i modelli di comportamento. Le feature vengono identificate ed estratte attraverso un processo euristico, creativo e trial & error, alla ricerca di ciò che è o interpretabile o maggiormente caratterizzante i comportamenti osservati.

Alla fine di questo importante processo sono state identificate un centinaio di

IMMAGINE 1



feature mediante le quali profilare i cookie, il loro comportamento di navigazione (fasce orarie), l'approccio al portale (aree visitate), gli interessi (basati sulle keyword incontrate nel corso della navigazione), il rapporto con la tecnologia (a partire dalle informazioni sulla piattaforma tecnologica usata), la geo-localizzazione (a partire dagli indirizzi IP).

4. Output visivo e relativa prototipazione

A questo punto si sono affrontate le tre linee di progetto condivise con Sky e per ciascuna si è identificata una proposta differente.

A. Descrizione visiva della navigazione

Per descrivere la navigazione al sito dei visitatori, si è fatto ricorso alle Self-Organizing Map (o Mappe di Kohonen), che consentono di proiettare dati multidimensionali su una superficie a due dimensioni, e quindi visualizzabile. È lo stesso strumento illustrato nel primo numero di Datavalue, nell'ambito di un progetto svolto con IGPDecaux. Come ricorderemo queste mappe forniscono una topologia (una mappa) della distribuzione dei tipi di oggetto in esame (in questo caso i navigatori del sito). In questo caso sono state costruite per capire come si distribuiscono i navigatori del sito in base a tre prospettive:

1. Giorni della settimana.
2. Fasce orarie (ore lavorative, ore festive ecc.).
3. Aree tematiche del sito.

Per descrivere la navigazione al sito dei visitatori, si è fatto ricorso alle Self-Organizing Map (o Mappe di Kohonen), che consentono di proiettare dati multidimensionali su una superficie a due dimensioni, e quindi visualizzabile

Ciascun profilo di mix delle caratteristiche prese in esame in ognuno dei tre casi corrisponde ad un colore. Il colore pieno è usato per profili con la presenza di una sola caratteristica (es: i rossi dell'ultima mappa sono coloro che navigano solo sull'area calcio). Mentre colori mischiati e sfumati per i vari profili che presentano la compresenza di più caratteristiche in misure differenti. Questi profili sono disposti sul mosaico in modo da raggruppare profili vicini e preservare tutti i mix e le zone di confine (*vedi immagine1*).

Dalle mappe emergono alcune evidenze: il comportamento degli utenti non mostra pattern significativi rispetto ai giorni della settimana nei quali si naviga, rivela alcuni pattern più marcati in termini di fasce orarie e pattern decisamente evidenti rispetto alle aree tematiche del sito. Si vede come alcune aree tematiche siano distinte e senza sovrapposizioni (per esempio Cinema e News), mentre altre siano contigue e in parte sovrapposte fino a sfumare una nell'altra (come Calcio e Assistenza). Ciò significa che mentre non vi è un numero significativo di visitatori dell'area Cinema che navigano anche l'area News, vi sono visitatori dell'area Calcio che visitano l'area di Assistenza.

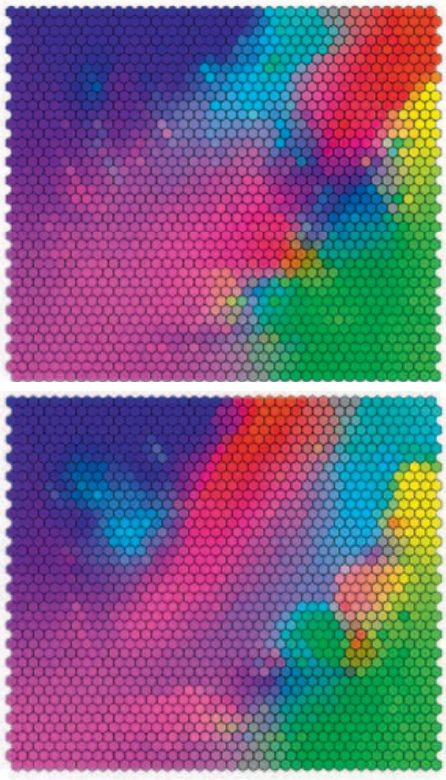


IMMAGINE 2

DAL CONFRONTO DELLE DUE MAPPE (RIPORTATE NELL'IMMAGINE 2) EMERGE IMMEDIATAMENTE LO SPOSTAMENTO DELL'AREA ROSSA (CALCIO) VERSO L'AREA BLU (NEWS) E INFATTI LA SECONDA MAPPA È RELATIVA ALLA FINE DEL CAMPIONATO E ALL'INIZIO DEL CALCIO MERCATO: GLI APPASSIONATI DI CALCIO FREQUENTANO PIÙ ASSIDUAMENTE L'AREA NEWS CHE CONTERRÀ MOLTE NOTIZIE DI LORO INTERESSE.
PER AVERE UN'IDEA CONCRETA DI COME I VISITATORI OCCUPINO LA GEOGRAFIA DEGLI INTERESSI (CHE NON RAPPRESENTA UNA METRICA MA INDICA SOLO LA DISTRIBUZIONE DEI POSSIBILI PROFILI), IL CAMPIONE USATO PER L'ANALISI È STATO PROIETTATO SULLA MAPPA: OGNI COOKIE, CORRISPONDENTE A UN PUNTINO, È STATO COLLOCATO NELLA CELLA CORRISPONDENTE AL PROPRIO PROFILO DI VISITA DELLE AREE DEL SITO (*vedi immagine3*).

SI VEDE CON IMMEDIATEZZA COME MOLTA PARTE DEI NAVIGATORI TENDA AD AVERE INTERESSI MONOTEMATICI, ANDANDO A OCCUPARE ZONE DELLA MAPPA CHE CORRISPONDONO A SINGOLE AREE DEL SITO. VI SONO PERÒ ANCHE UTENTI CON INTERESSI "MIX" E DIRETTRICI DI "MIXING" TRA AREE TEMATICHE. LE GEOGRAFIE RIPRODOTTE DALLE MAPPE POSSONO ESSERE UTILIZZATE COME SFONDO PER LA PROIEZIONE DI CLASSI E SEGMENTI DI VISITATORI, PER ESEMPIO: I NAVIGATORI DEL LUNEDÌ, GLI UTENTI MOBILE (O ANCHE QUELLI CHE USANO UN DEVICE APPLE), GLI UTENTI DEI GROSSI CENTRI ABITATI O GLI UTENTI DEL CENTRO ITALIA ECC. E' ANCHE POSSIBILE OSSERVARE LO SPOSTAMENTO DEL COMPORTAMENTO DEI VISITATORI NEL TEMPO, OSSERVANDONE LE DINAMICHE SULLO SFONDO DELLA GEOGRAFIA DI RIFERIMENTO.

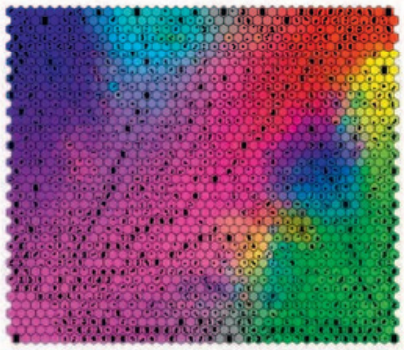


IMMAGINE 3

Molto interessante è anche osservare come la mappa cambia nel tempo (da un anno all'altro, per comprendere come cambia la propria utenza) o come muta al cambiare delle stagioni. Vediamo qui a sinistra come esempio le mappe relative a due stagioni differenti (*vedi nell'Immagine 2*).

B. Criteri di ottimizzazione dell'auto-promozione

Quanto realizzato mappando gli utenti può essere utilizzato per identificare dove (nella mappa delle aree) o quando (nella mappa delle fasce orarie) navigano principalmente coloro che acquistano i diversi tipi di pacchetto. In questo modo si possono impostare azioni di auto-pubblicità in modo più efficace, uscendo anche dallo schema "pubblicità calcio sull'area calcio", "pubblicità cinema sull'area cinema" e così via. Sul tema dell'auto-pubblicità si è poi deciso di sviluppare un modello più raffinato, che apre a diversi tipi di applicazioni, alcune più semplici e immediate e altre più efficienti, ma che necessitano di adeguamenti tecnologici.
È stato stimato statisticamente quanto ciascuna delle circa cento feature estratte dai cookie impatti sulla probabilità che un cookie acquisti "un certo" prodotto Sky. Questo ha portato a stimare per ognuna della feature uno score di "impatto" per ogni pacchetto. Vediamo di sotto un estratto di una di queste tabelle che evidenzia solo due righe: quella relativa alla feature che misura la percentuale di pagine viste nella pausa pranzo in giorni lavorativi e quella relativa alla percentuale di pagine di previsione weekend viste rispetto al totale (*vedi Tabella 1*).

Questa tabella può essere letta in orizzontale, per decidere in modo puntuale quali pacchetti promuovere indipendentemente da chi la sta vedendo

	calcio	sport	cinema	famiglia	HD
pranzo feriale	0,46	0,39	-0,14	0,52	0,38
previsioni weekend	0,66	2,76	0,00	0,00	0,00

Tabella 1: Impatto fra effetti dei cookie e acquisto pacchetti.

Appena un utente metterà un prodotto nel carrello, il suo profilo verrà passato al modello statistico, ottenendo in risposta gli score di propensione all'acquisto degli altri 4 pacchetti Sky

(ad esempio nelle diverse aree del sito o nei diversi momenti della giornata). Nell'esempio riportato nella tabella emerge come, nel periodo analizzato, in pausa pranzo navigano soprattutto coloro che sono propensi ad acquistare il pacchetto calcio e il pacchetto famiglia, mentre quelli che vedono le previsioni del weekend sono poi quelli che più probabilmente acquistano il pacchetto sport.
Ma la tabella può anche essere letta in verticale, in modo da proporre al singolo cookie il pacchetto che più probabilmente potrebbe acquistare, in base al suo particolare e unico profilo di feature. Questo necessita di un sistema di erogazione in grado di ricevere dal modello le probabilità di interesse per ogni singolo pacchetto del cookie che vede qualsiasi pagina del sito. Si può andare anche oltre e analizzare tutti questi dati nei diversi anni, per cogliere le dinamiche di comportamento dei visitatori e reagire ai cambiamenti, o nelle varie stagioni per calibrare meglio le azioni, contestualizzandole al periodo corrente.



C. Strumenti di ottimizzazione dell'up-selling

Individuare la proposta di upselling giusta per un certo cliente implica agire a livello disaggregato, sul singolo utente. L'idea di base è stimare, con un modello analogo a quello precedente, la probabilità che un certo cookie, con un certo profilo e che stia mettendo nel carrello un certo prodotto, acquisti anche un altro prodotto dell'offerta Sky, e quindi proporgli di aggiungere il prodotto con la stima di probabilità più alta nel momento in cui il carrello viene aperto. Appena un utente metterà un prodotto nel carrello, il suo profilo verrà passato al modello statistico, ottenendo in risposta gli score di propensione all'acquisto degli altri 4 pacchetti Sky.

Anche in questo caso, è utile considerare modelli stagionali, perché le evidenze raccolte nell'attività pilota indicano l'esistenza di comportamenti e attitudini differenti nelle diverse stagioni e legati alla periodicità di eventi e fenomeni sportivi.

Applicazioni ed evoluzioni

Come si è appena visto e come spesso capita, valutando tanti approcci, in realtà si è arrivati a più di un risultato interessante e si è deciso di presentarli tutti. Il progetto pilota ha soddisfatto le esigenze del Committente, fornendo viste e "insights" giudicati interessanti e ricchi di chiavi di lettura.
Il passo successivo sarà quello di implementare un processo ingegnerizzato che generi sistematicamente gli output illustrati in precedenza, in modo da monitorare l'evoluzione dei comportamenti dei visitatori e intercettare tempestivamente segnali di cambiamento. Contemporaneamente, sono già state immaginate possibili evoluzioni, per affrontare temi legati all'incremento del volume di navigazione sul portale (attraverso l'implementazione di un motore di suggerimento intelligente), all'aumento del tasso di "conversion" dei prospect provenienti da campagne esterne (integrandosi con i dati presenti sulla DMP), alla riduzione del tasso di churn (attraverso un'analisi e una profilazione più raffinata dei comportamenti). Temi che, se affrontati con metodologie appropriate e la giusta componente di pensiero laterale, porteranno a nuovi e interessanti risultati.

Segmentare e visualizzare un MERCATO COMPLESSO

IGPDecaux ha avviato un progetto di segmentazione del mercato, in grado di fornire un quadro dinamico della sua struttura e delle varie tipologie di comportamento suddivise per canali di comunicazione. Con un obiettivo di fondo: reagire tempestivamente alle fluttuazioni della domanda



di Marco Fattore, Silvio Gerli
(marco.fattore@unimib.it)
(silvio.gerli@smartstat.it)

datavalue

Il tema della segmentazione (del mercato, dei clienti, dei prodotti...) occupa una parte decisamente rilevante dei processi di business intelligence interni alle direzioni marketing. Un tema noto, per il quale esistono metodologie, modelli e strumenti implementati nelle principali piattaforme analitiche e quotidianamente utilizzati dai business analyst. Ciononostante, un processo che è molto più sottile, critico e strategico di quanto si possa supporre. Il punto è il seguente: pensare il mercato come costituito da segmenti ben definibili, stabili nel tempo e a cui corrispondono comportamenti chiari e ripetuti è incoerente con un'economia così dinamica e poco prevedibile come quella attuale. La quantità di variabili che influenzano i criteri di scelta di consumatori e investitori, la loro interdipendenza, la continua innovazione dell'offerta, le condizioni strutturali dell'economia e le stesse dinamiche sociali generano comportamenti economici non riducibili a schemi semplici e fissi. Due sono le parole chiave: **complessità e dinamicità**. Gli attori economici si **muovono** in un contesto **complesso, multidimensionale, articolato e dinamico** e i loro comportamenti non possono essere descritti prescindendo da questo. Segmentare il mercato significa quindi catturare e riprodurre queste dinamiche, pur semplificandole, con l'obiettivo di **cogliere tempestivamente le direzioni di cambiamento**. Un processo di segmentazione deve quindi essere costruito come un **processo continuo di osservazione del mercato** e di confronto (nel tempo, sul territorio, fra strati sociali, fra settori...) dei comportamenti dei soggetti economici.

Questo punto di vista ha dei riflessi operativi **sul design dei processi di segmentazione, sulla scelta delle tecniche di analisi dei dati e sulle modalità di comunicazione e interpretazione dei risultati.** Affrontiamo questi aspetti discutendo un recente caso aziendale, in cui è stato progettato e reso operativo un processo di segmentazione, relativo ai comportamenti di investimento nel mercato della comunicazione pubblicitaria.

Il **contesto** del **mercato** pubblicitario e il **comportamento** degli investitori in comunicazione sono **radicalmente mutati** nel corso degli ultimi dieci anni, con un'accelerazione a cavallo dello scoppio della crisi economica. Un mercato che, almeno in termini di volume, aveva i propri punti di riferimento in tre grandi mezzi di comunicazione – televisione, quotidiani e periodici – vede cambiare i criteri di associazione tra canali, settori economici e prodotti comunicati. Emergono nuovi mezzi – internet e televisioni satellitari su tutti - che ampliano e articolano l'offerta di comunicazione. Prodotti tipicamente associati ai mezzi tradizionali migrano verso nuovi canali. Gli investitori differenziano i propri profili di investimento, con media-mix alternativi. In sintesi, il mercato della comunicazione è sottoposto a una profonda **dinamica strutturale**: mutano i volumi di investimento e i mix di allocazione sui differenti mezzi e tutto ciò cambia in funzione del settore economico dell'investitore.

Il **problema** da cui ha preso le mosse il progetto non è principalmente analitico, ma di **governance aziendale**. L'avvento della crisi economica e le dinamiche proprie del mondo della comunicazione pubblicitaria hanno repentinamente trasformato un mercato "stabile e sicuro", in un mercato "da interpretare e da andare a prendere". Precondizione: conoscere il mercato e coglierne le dinamiche. Il Committente, IGPDecaux, ha quindi deciso di costruire un processo di segmentazione, che fornisse un quadro dinamico della struttura del mercato, del comportamento degli investitori e dei flussi di investimento fra mezzi di comunicazione. Il tutto reso poi disponibile sotto forma di un'applicazione software che consentisse ai business analyst di interagire visualmente con i risultati, selezionando e confrontando i dati secondo prospettive temporali ed economiche diverse.

I dati utilizzati sono costituiti da un flusso mensile di rilevazioni statistiche contenenti i **volumi di investimento** per ciascuna azienda censita (complessivamente, più di 26.000 investitori) sui mezzi di comunicazione considerati (per un totale di 14 canali differenti). Per ricostruire la struttura del mercato, l'informazione principale è tuttavia rappresentata dai profili di investimento delle aziende, cioè dalle quote di investimento in comunicazione che ciascun investitore alloca sui diversi mezzi; dunque la base dati di input viene completata con la **ricostruzione della composizione annuale dell'investimento**, per ciascuna azienda. La scelta dell'anno come unità temporale è poi dettata dalla necessità di eliminare fluttuazioni stagionali che sarebbero di interesse relativo, per la ricostruzione generale delle dinamiche del mercato. In definitiva, la base dati utilizzata è costituita da un panel temporale di vettori, ciascuno dei quali associato a un'azienda e contenente la struttura percentuale del suo investimento in comunicazione, per ognuno degli anni considerati (nel caso in esame, a partire dal 2009). A questi, vanno aggiunte le **informazioni sul settore economico** di appartenenza delle aziende e le relative suddivisioni in sotto-settori, che precisano la tipologia di prodotto o di servizio comunicato.

Il **processo di segmentazione** è stato costruito a partire da un'evidenza che i dati hanno fatto emergere ai primi tentativi di analisi: **il mercato degli investimenti pubblicitari non può essere descritto come composto da pochi tipi di comportamento "paradigmatico", ben identificabili e distinti.** Certo, si possono identificare attitudini generali (per esempio, molte aziende effettuano ancora comunicazione commerciale utilizzando un solo tipo di canale) e certi settori economici sono più affini a certi mezzi, che ad altri. Ma il comportamento degli investitori è decisamente **sfumato**: non esistono cluster veri e propri, ma solo **pattern di investimento**, che caratterizzano e diversificano i settori economici. Le aziende, cioè, non si concentrano su pochi tipi di media-mix, ma occupano "intere zone di comportamento", caratterizzate dalla presenza di profili di investimento progressivamente differenziati, per tipologia di canale e ripartizione delle quote investite. Questo ha un riflesso immediato sul processo statistico di analisi dei dati: **gli strumenti di clustering "classici" sono inadeguati,** non sono in grado di riprodurre con fedeltà i pattern esistenti e producono viste distorte del mercato.

I dati utilizzati sono costituiti da un flusso mensile di rilevazioni statistiche contenenti i volumi di investimento per ciascuna azienda censita sui diversi mezzi di comunicazione considerati

CORNER

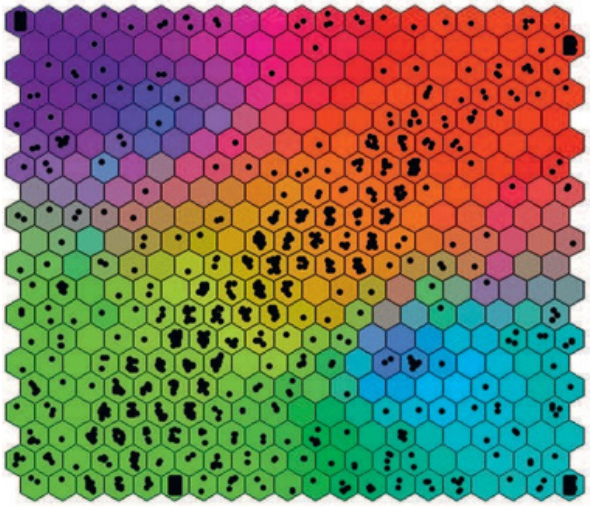
Rmette a disposizione molte funzioni per l'esecuzione di cluster analysis,in diversi package scaricabili direttamente dal sito cran.r-project.org.
Ma anche la configurazione base del linguaggio, quella che viene resa operativa quando R viene installato, contiene procedure per effettuare cluster analysis gerarchiche (funzione *hclust*) e cluster analysis non gerarchiche (funzione *kmeans*).
Queste consentono di effettuare analisi di raggruppamento classiche con pochissime righe di codice, generando i grafici e tutte le diagnostiche necessarie a interpretare i risultati e valutarne l'adeguatezza.
Le mappe di Kohonen sono invece implementate nel pacchetto kohonen, attraverso la funzione *som* (acronimo di "Self-organizing maps").
Anche in questo caso, il complesso algoritmo di calcolo è racchiuso in una semplice istruzione che genera tutte le informazioni utili alla rappresentazione grafica dei risultati e alla loro successiva interpretazione. Lo sviluppo di applicazioni web è poi possibile utilizzando il pacchetto shiny.
Tutti i dettagli sulle istruzioni e sui pacchetti di R sono rintracciabili sul sito cran.r-project.org. Comunque, digitando *?hclust* (o qualunque altra istruzione) nell'editor di R, si accede direttamente all'help. Ed essendo R un linguaggio open source, digitando il nome di un'istruzione se ne può ispezionare anche il codice.

Normalmente, infatti, i processi di segmentazione vengono impostati utilizzando metodologie di raggruppamento che cercano di individuare cluster di soggetti compatti, omogenei e ben distinguibili. Il metodo più utilizzato, anche perché di semplice esecuzione e disponibile in molte delle piattaforme di business intelligence, è probabilmente quello noto come "metodo delle k-medie": il business analyst decide il numero di gruppi in cui desidera siano suddivisi i record del proprio dataset e la procedura ripartisce le unità statistiche, individuando alcuni profili di riferimento ("prototipi"), intorno ai quali effettuare l'aggregazione. Verificando l'omogeneità interna dei gruppi e la loro separazione, ed eventualmente modificando il numero di gruppi richiesto, è poi possibile affinare la segmentazione, cercando di raggiungere una sintesi dei dati che sia ragionevole e utile. Ma nel mercato della comunicazione sono proprio i profili "prototipali" ad essere assenti, rendendo pressoché inutilizzabile un approccio di questo tipo. Per rivelare i pattern nascosti nei dati, si è fatto quindi ricorso a uno strumento di analisi alternativo: le cosiddette **"Mappe di Kohonen" o "Self-organizing map"**. Si tratta di algoritmi computazionalmente complessi, ma ormai implementati nelle principali piattaforme analitiche. Il loro utilizzo nell'ambito del marketing è piuttosto diffuso, almeno nei contesti più innovativi, per la loro capacità di riprodurre la **"geografia" del mercato**, producendo **visualizzazioni** che spesso sono **gradevoli e facilmente interpretabili**. Esse infatti generano cluster di unità statistiche in modo parzialmente simile al metodo delle k-medie, ma li associano ai punti di una mappa bidimensionale, in modo che questi risultino alla fine disposti come le tessere di un mosaico, in cui cluster di individui simili sono tendenzialmente disposti in posizioni contigue. Sovrapponendo ai cluster opportuni codici di colore, si genera infine una visualizzazione "parlante", che fa **emergere la struttura dei dati**, cioè, nel caso in esame, dei pattern di media-mix sfruttati dagli investitori. La lettura delle mappe è quindi orientata all'interpretazione della loro struttura generale e non alla analisi precisa delle caratteristiche dei singoli cluster. Applicando questi algoritmi, con l'inevitabile processo "trial and error" necessario ad affinare i risultati, sono state generate diverse mappe visuali, che riproducono la geografia del mercato degli investimenti. Le mappe vengono poi utilizzate come "sfondo" per l'interpretazione dei comportamenti degli investitori: si possono così osservare le traiettorie temporali di singole aziende, o confrontare le strutture di investimento di settori economici diversi, cogliendone le affinità o le differenze (come esemplificato nella figura che accompagna l'articolo) o anche individuare gruppi di aziende di interesse, per esempio in base al volume totale investito in comunicazione o a qualunque altro parametro disponibile. Il tutto, con l'obiettivo di "vedere quello che accade", comprenderlo, interpretarlo e migliorare la propria capacità di decisione, di adattamento al contesto e di immaginazione di azioni e percorsi innovativi.

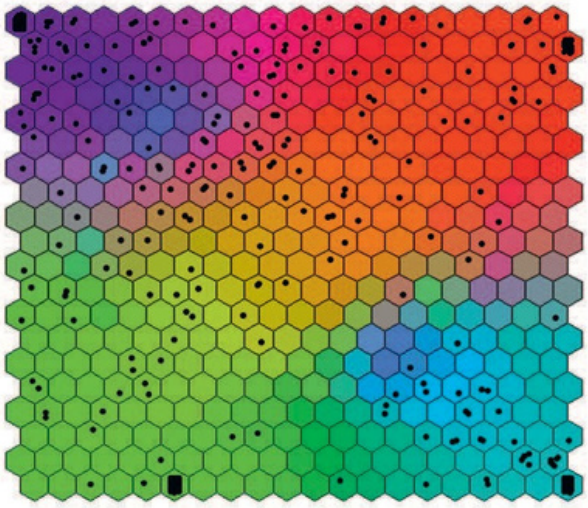
L'applicazione software che rende operativo il processo di segmentazione consente quindi di **generare, visualizzare** e **confrontare mappe** del mercato differenti, con una semplice **interfaccia web**. È possibile selezionare i mezzi di comunicazione rispetto ai quali effettuare l'analisi, scegliere gli anni di interesse, proiettare graficamente sulla mappa le aziende per individuarne pattern di comportamento, scegliendo i settori e i sotto-settori di interesse, associare a ogni zona della mappa i valori di altre variabili e dunque visualizzare i cambiamenti strutturali del mercato. Delle singole aziende è poi possibile conoscere le informazioni di dettaglio, interagendo con la visualizzazione. L'applicazione diventa così un ambiente per l'osservazione del mercato da una prospettiva manageriale. Essa non si limita

LE DUE MAPPE DI KOHONEN RAPPRESENTANO LA STRUTTURA DEL MERCATO DEGLI INVESTIMENTI PUBBLICITARI, relativamente a quattro canali di comunicazione: Periodici (in verde), Quotidiani (in rosso), Affissioni e Comunicazione Esterna (in azzurro) e Internet (in violetto). Il colore dei singoli tasselli corrisponde al mix di investimento nei quattro mezzi considerati. Tasselli "verdi" indicano investimenti pubblicitari concentrati sui periodici; tasselli che sfumano tra il verde e l'azzurro indicano media-mix che coinvolgono una quota di investimento in periodici e una in affissioni e così via, per le varie gradazioni cromatiche osservabili. I punti scuri sovrapposti alle mappe rappresentano aziende investitrici del settore dell'Abbigliamento (sinistra) e del settore dell'Automobile (destra). Le mappe mostrano un ordine cromatico (auto-organizzazione) che rivela la struttura dei dati e che non è stato imposto dall'analista, ma fatto emergere dall'algoritmo stesso. Osservando lo sfondo delle mappe, si nota come nel mercato siano presenti alcuni tipi di media-mix (relativamente ai mezzi qui considerati) e non altri e come, in particolare, sia ancora presente un asse Quotidiani-Periodici (corrispondente alla zona rossa, a quella verde e a quella intermedia alle due, che è quindi sfumata su toni arancioni). Sono presenti "zone di passaggio" anche fra gli altri mezzi, ma in misura meno progressiva, perché nei dati sono meno presenti investitori che utilizzano i corrispondenti media-mix. Proiettando i punti corrispondenti alle aziende, sullo sfondo delle mappe, i pattern di comportamento dei due settori si rivelano profondamente diversi. Il mercato dell'Abbigliamento vede la presenza di molti operatori, i cui comportamenti di investimento sfumano soprattutto lungo l'asse Quotidiani-Periodici, pur essendo presenti anche mix che coinvolgono Periodici e Affissioni, nonché Periodici e Internet. Vi sono poi aziende che investono in un solo canale di comunicazione. Il mercato dell'Automobile, invece, oltre a comprendere meno attori, ha un pattern di investimento molto più polarizzato sui singoli canali: l'investimento su un solo mezzo di comunicazione è prevalente, anche se sono presenti media-mix lungo altri assi. L'output è estratto dalla piattaforma descritta nel testo dell'articolo, attraverso la quale è poi possibile accedere alle informazioni sulle singole aziende proiettate sulla mappa.

Abbigliamento



Automobile



a mettere a disposizione i dati, fornendo strumenti per la loro esplorazione, ma trasforma i dati, generando nuovi layer informativi, più significativi per i decisori. Lo sviluppo software è stato effettuato con R, un linguaggio di programmazione open-source e gratuito (cran.r-project.org), dedicato all'implementazione di analisi e processi statistici, ormai strumento di riferimento nella ricerca universitaria e in via di diffusione anche nelle istituzioni e nelle aziende. L'adozione di R nel progetto è stata motivata da due ragioni: l'economicità e il fatto che i suoi moduli mettono a disposizione una quantità estremamente ampia di funzioni e di modelli statistici, tra i quali gli algoritmi relativi a vari tipi di cluster analysis e, naturalmente, alle mappe di Kohonen. Questi due aspetti, fra loro combinati, hanno consentito di affrontare lo sviluppo del processo di segmentazione, attraverso un percorso "agile", svolto per approssimazioni successive, in un continuo confronto con il Committente; il tutto mantenendo limitato il costo dell'attività.

Il risultato finale dell'intero progetto è dunque un processo continuo di ricostruzione dell'immagine del mercato e della dinamica degli investimenti pubblicitari, che consente al management di rendersi conto tempestivamente di ciò che sta accadendo nell'ecosistema in cui l'azienda è immersa. Non una semplice analisi dei dati, quindi, ma una strutturazione organica dell'informazione. **La trasformazione di input complessi e multidimensionali** in architetture di informazione significativa è uno dei tratti propri e caratteristici dei progetti di "data science". Progetti che infatti non si riducono ad attività tecnologiche, statistiche o di data visualization, ma che integrano queste dimensioni in soluzioni e servizi **guidati dai bisogni del management** e misurati in termini di impatto sulla sua capacità di comprensione e di azione.

CLUSTER ANALYSIS

Con il termine di Cluster Analysis (o Classificazione non Supervisionata) si intende quella parte dell'analisi dei dati multidimensionali dedicata all'identificazione di gruppi all'interno di una popolazione statistica.

Ciò che rende la Cluster Analysis un processo complesso e, per certi versi, sottile è che la nozione stessa di "gruppo" è fortemente ambigua. Che cos'è un gruppo di unità statistiche e che cosa lo rende identificabile come tale? Un gruppo di aziende potrebbe essere pensato, come spesso si fa, come un insieme di unità molto simili fra loro rispetto alle variabili statistiche rilevate (per esempio, dimensione, fatturato, utile...). In questo caso, la ripartizione della popolazione viene effettuata utilizzando metodi che aggregano unità "per similitudine". Sono i classici metodi di clusterizzazione gerarchica e di analisi delle k-medie. In entrambi i casi, si cerca generalmente di costruire gruppi contenenti unità fra loro simili, cercando un compromesso tra l'omogeneità interna dei cluster e il loro numero (che dovrebbe essere sufficientemente basso da produrre una vera semplificazione dei dati). Meno diffusi, ma potenzialmente utili, sono gli algoritmi "divisivi", nei quali l'insieme delle unità statistiche viene progressivamente suddiviso in gruppi, per separazione delle unità più dissimili. Ma nella realtà, spesso i cluster "compatti" non sono rintracciabili. Sono invece presenti pattern di comportamento sfumati, in cui le unità statistiche si differenziano progressivamente, senza soluzione di continuità, modulando tra profili alternativi. Le tecniche di clustering classiche non sono in grado di riprodurre efficacemente queste situazioni. Più adatte sono le cosiddette "Mappe di Kohonen" o "Self-organizing map" (o "Som"), algoritmi progettati esattamente per cogliere "la forma dei dati" e riprodurla visivamente su mappe bidimensionali. Tecnicamente, si tratta di reti neurali che si strutturano in base ai dati in input e ne riproducono "l'immagine", approssimandola su un piano, e quindi rendendola visibile. Rispetto ad altri algoritmi di visualizzazione di dati multidimensionali (per esempio il "multidimensional scaling", piuttosto utilizzato nel marketing), le mappe di Kohonen sono più efficienti e capaci di adattarsi a dati più complessi e articolati. La loro facilità di implementazione le rende uno strumento estremamente interessante, in molti campi di interesse aziendale (per esempio, esse sono state utilizzate per individuare clienti a rischio di "churn" nel mercato Telco, per analizzare le dinamiche dei Tweet nella sentiment analysis, per la segmentazione del mercato, come nell'esempio riportato nel testo dell'articolo, per classificare prodotti, per effettuare analisi socio-demografiche...).