PROGETTI CONCLUSI II CALL	
Titolo progetto	Parallelizzazione dell'algoritmo di Metropolis-Hastings per la stima per intervallo della dimensione della popolazione
Descrizione	Si vogliono ottenere delle stime per intervallo della dimensione della popolazione disaggregata per anno (dal 2006 al 2014), area geografica (19 regioni e le due province autonome di Trento e Bolzano), sesso ed età (90 classi annuali ed una classe residuale di 90 e più anni) a partire dalle serie storiche demografiche pubblicate dall'Istituto (dal 2016 dal Registro Base degli Individui). Per ognuna delle celle, identificate dalle 4 coordinate dette, si vuole una stima dei conteggi di popolazione che tenga conto sia della relazione tra popolazione nell'anno corrente con quella dell'anno precedente, i conteggi di nati, decessi e migrazioni (equazione della popolazione), sia della relazione tra dati da archivio osservati affetti da errore di copertura e valori "veri" non osservabili privi di errore (modelli di errore di misura). La formulazione di un modello di stima che tenga conto del vincolo tra le variabili in analisi rappresentato dell'equazione della popolazione e dai modelli d'errore degli archivi, è un problema metodologico con un elevato grado di complessità. Bryant e Graham propongono un modello bayesiano in grado di includere entrambi gli aspetti. Per definire un modello bayesiano occorre specificare le distribuzioni di probabilità di tutte le quantità in studio, che sono quindi considerate variabili aleatorie, introducendo le eventuali informazioni a priori, come ed esempio il campo di variazione delle diverse grandezze, tramite i parametri delle distribuzioni. Ricorrendo poi al teorema di Bayes si aggiornano le distribuzioni a priori tramite la verosimiglianza, ottenendo le distribuzioni a posteriori di tutte le variabili del problema. Dopo aver adattato il modello proposto di Bryant e Graham alla situazione italiana, specificando opportunamente la struttura delle relazioni di dipendenza probabilistica tra le variabili, si è passati a considerare il problema della stima. In questo caso non è possibile procedere alla stima del modello bayesiano per via analitica a causa della sua complessità, è però possibile utilizzare tecniche simulative di stima come MCMC (Markov Chain Monte Carlo). L'algoritmo di Metropolis- Hastings è, nel caso di variabili latenti discrete, uno dei metodi più usati. L'utilizzo di tecniche simulative richiede la determinazione dei valori iniziali delle variabili in studio da cui far partire processi aleatori markoviani (dette Catene) che, arrivate a convergenza, estrarranno valori dalla distribuzione obiettivo. Nel caso di modelli complessi, tali valori se presi troppo "lontani" dai valori di convergenza, determinano tempi molto lunghi di "avvicinamento". Occorre quindi fissare dei valori iniziali "ragionevoli" cioè entro il campo di variazione atteso per la variabile, in base alla conoscenza del fenomeno. La formulazione completa del modello prevede la stima dei conteggi latenti di nascite, decessi, immigrazioni ed emigrazioni interne ed esterne per un totale di 6 componenti articolate in 30576 celle (21 aree geografiche x 2 sessi x 91 classi di età x 8 anni) ognuna per le quali fissare valori iniziali, oltre ai parametri delle distribuzioni e a quelli relativi ai modelli d'errore (i parametri di copertura sono anch'essi 30576 per le 6 componenti, più i 30576 relativi alle celle dei conteggi di popolazione). La dimensione computazionale del modello suggerisce di procedere per passi alla sua stima. Il primo passo consiste nello specificare il modello d'errore solo per una componente, assumendo quindi che le altre 5 non siano affette da errore di copertura. Si è dunque deciso di assumere solo per i conteggi dei decessi, la presenza di un errore di copertura.

	Il modello è stato trascritto nel linguaggio del software WinBUGS. Quest'ultimo, infatti, permette di stimare i modelli bayesiani per via simulativa in modo sequenziale. I tempi di calcolo sono risultati molto lunghi (oltre i 6 giorni) perché la complessità e la dimensione del problema di stima richiede un numero elevato di iterazioni del processo prima di osservare la convergenza delle stime, per studiare la quale occorre lanciare 2 processi di stima da due diversi punti di inizio. Ogni iterazione, corrispondente all'accettazione di uno "spostamento" del processo verso i valori proposti dall'algoritmo di ricerca nello spazio dei parametri, richiede tempi relativamente lunghi come atteso dalla teoria. La fase successiva è stata la valutazione della fattibilità dell'introduzione della componente d'errore per le migrazioni. Il modello specificato da Bryant e Graham per le migrazioni non risulta adatto alla situazione italiana fortemente caratterizzata da variabilità tra regioni. Per tenere conto di questa eterogeneità occorre introdurre ulteriori parametri da stimare. A questo punto la scelta del software per la stima è diventata fondamentale e l'utilizzo del software precedentemente identificato è risultato non soddisfacente. Dopo una ricognizione nella letteratura, ci si è orientati verso l'utilizzo di librarie R che utilizzano TensorFlow di Google. Si sta procedendo con la traduzione del codice fin qui sviluppato in BUGS nel nuovo linguaggio scelto. La valutazione del nuovo approccio alla stima indicherà i possibili successivi passi.
Obiettivi	In un'ottica di controllo di qualità dei valori di conteggio della popolazione raccolti dall'Istat per aggiornare le proprie banche dati (dal 2016 il Registro Base degli Individui), si vuole fornire uno strumento in grado di valutare la coerenza tra l'informazione storica e il "nuovo" dato di conteggio della popolazione. Dal confronto tra i valori di conteggio forniti dagli archivi amministrativi (o dal registro) e le relative stime per intervallo, ottenute sulla base dei dati demografici storici e di ipotesi sulla dimensione dell'errore di misura degli archivi, si possono ricavare indicazioni sulla tenuta nel tempo di tali ipotesi e sulla dimensione attuale dell'errore di copertura. La rimodulazione degli assunti sulle caratteristiche dell'errore di copertura degli archivi permette, infatti, di valutare quali ipotesi siano più compatibili col dato osservato.
Metodologia	Il problema della stima dei conteggi di popolazione è formalizzato tramite un modello bayesiano a variabili latenti. La struttura latente del modello è quella relativa ai "veri" valori di conteggio per le diverse poste dell'equazione di bilancio della popolazione: conteggi di nascite, decessi e migrazioni vengono descritte indipendentemente tramite modelli Poisson-Gamma. In particolare, si modella il valore atteso dei conteggi latenti tramite modelli di regressione che considerano età, sesso, regione e anno come potenziali regressori. Una volta introdotte le quantità latenti nell'equazione della popolazione e aggiunto il totale della popolazione dell'anno precedente, l'equazione fornisce il "vero" valore di conteggio per l'anno d'interesse. Sia le quantità latenti a destra dell'equazione che il totale di popolazione a sinistra, vengono messi in relazione ognuno coi relativi valori osservati tramite un modello di errore che caratterizza l'associazione tra valore osservato e valore latente tramite un parametro interpretabile come il livello di copertura dell'archivio dei valori "veri". La misura della copertura è descritta tramite una distribuzione di tipo Gamma sui cui parametri vengono fatti assunti distribuzionali debolmente informativi. È a questo livello che si introducono le informazioni ed assunti circa l'entità dell'errore di copertura degli archivi.

Risultati ottenuti specificare l'impatto sulla produzione statistica	A conclusione della prima fase del progetto, sono state ottenute le distribuzioni a posteriori per i conteggi di popolazione relativi alle 3822 celle (21 aree geografiche x 2 sessi x 91 classi di età) del 2014, avendo assunto un valore atteso per l'errore di copertura dell'85% sia per i conteggi di popolazione che per i conteggi dei decessi. La seconda fase ha permesso la formulazione di un modello d'errore per le migrazioni coerente con la situazione italiana e l'individuazione di alcune librarie di R in grado di stimarlo.
Membri del team **Nome cognome e indirizzo e mail**	Simona Toti, Rosa Maria Lipsi, Sara Giavante, Stefano Daddi.