Likelihood-Free Parameter Estimation for Dynamic Queueing Networks: The Case of the Immigration Queue at an International Airport

Fabrizio Ruggeri

Istituto di Matematica Applicata e Tecnologie Informatiche

Consiglio Nazionale delle Ricerche

Via Bassini 15, I-20133, Milano, Italy, European Union

fabrizio@mi.imati.cnr.it

web.mi.imati.cnr.it/fabrizio/

Joint work with

- Anthony Ebert, Antonietta Mira
  Universitá della Svizzera Italiana, Lugano, Switzerland
- Kerrie Mengersen, Paul Wu
  Queensland University of Technology, Brisbane, Australia
- Ritabrata Dutta
  University of Warwick, United Kingdom

OUTLINE OF THE TALK

- Motivating example: queues at immigration control in airports
- Queues and dynamic queueing networks
- R package `queuecomputer` to simulate queues
- Approximate Bayesian Computation (ABC)
- Application of the method to immigration control
- Critical aspects
- Future and possible research

MOTIVATION OF THE WORK

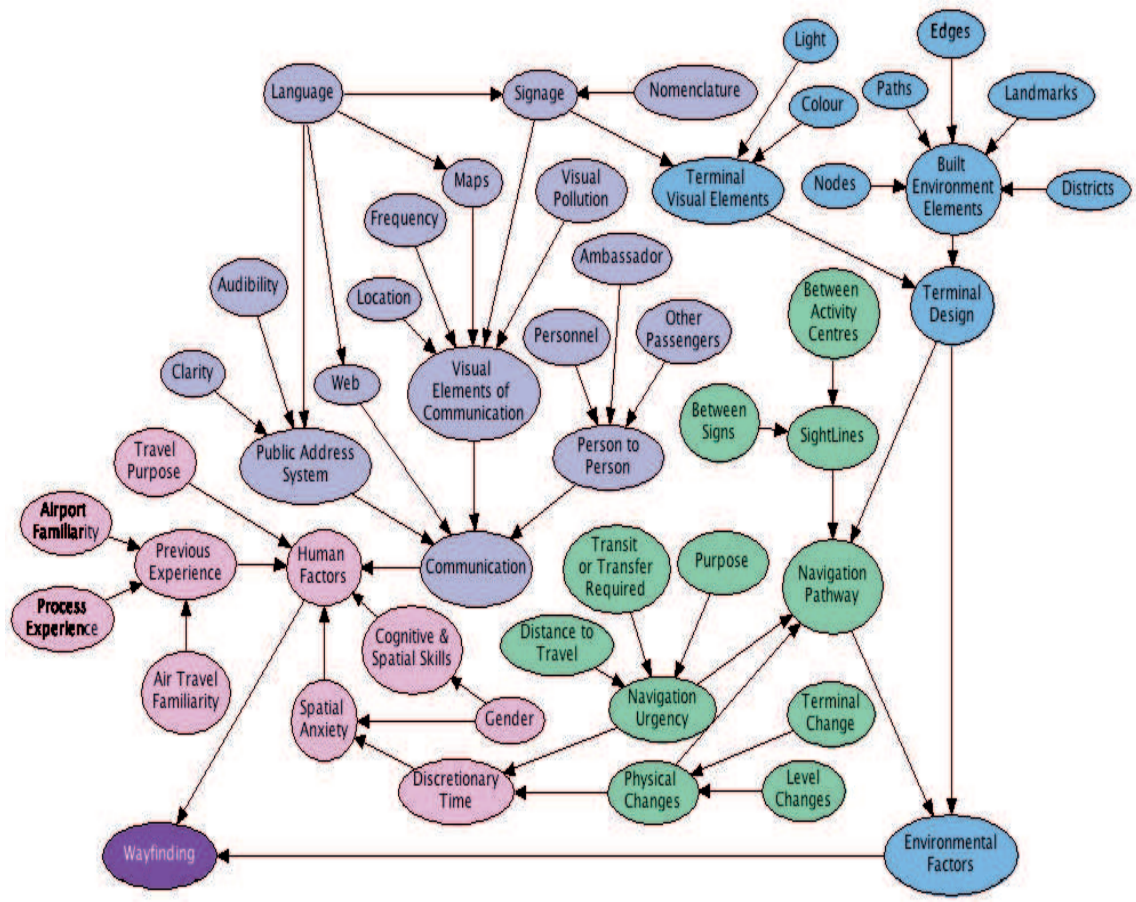
- Demand for air transport rapidly increasing, e.g., per year,
  - from 98.1 (2009) to 124.9 (2015) million passengers within Australian capital cities
  - worldwide from 2.7 billion (2011) to 4 billion (2017) and predicted 8.2 billion (2037)
- More comprehensive security and immigration screening requirements
- $\Rightarrow$ Critical consequences
  - increase of passengers' congestion within airports
  - more flight delays
  - more pressure on existing infrastructure and operations
  - capacity constraints on future growth
- Interest in smoother and quicker movement of the passengers within terminals

EARLIER WORK AT QUT

Wayfinding Bayesian Network Model (WBNM)

- Airports of the Future project
- WBNM developed to investigate the factors that influence effective wayfinding in airports
- Human and environmental factors considered in WBNM
- WBNM: Bayesian network (BN) constructed through focus groups, wayfinding literature and online surveys
- Conditional probabilities built out of 99 online surveys (*experts' opinions*)
  - one BN based on prior pooling of opinions at each node
  - merge of 99 BNs built for each expert (posterior pooling)
  - Measurement Error Approach, with opinions at each node considered as noisy observations of the *true* probability value

WAYFINDING BAYESIAN NETWORK MODEL



IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

- Passenger facilitation as a series of processes designed to make the journey through an airport as smooth as possible

- Passenger facilitation concern for duty and operations managers
  $\Rightarrow$ need to predict and avoid congestion

- Interest in arriving passengers at international airport terminals

- Multiple stages traversed by passengers
  - airside concourse
  - immigration
  - baggage collection
  - customs
  - landside concourse

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

- Interest in modelling passenger movements through the terminal
- Passenger terminal simulations developed for four main purposes
  - capacity planning
  - operations planning
  - security policy and planning
  - airport performance measurement
- Motivating case study: operational planning of immigration control
  - conflicting consequences: cost and availability of officers vs. length of queue
  - complexity due to number of passengers and incoming flights and other factors (distance from arrival gates, waiting on board before disembarking, stop at stores and/or restrooms, individual walking speed, etc.)
  - $\Rightarrow$ need for *quick* online inference for day-to-day (and real time) decisions on number of open booths

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

- Simulation studies and/or models for many systems within an international airport passenger terminal
  - flight delays
  - gate assignment
  - runway queueing
  - performance evaluation
  - effect of global airport network on global epidemics
  - security systems
  - passenger walking speed
  - passenger wayfinding (WBNM)

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

- Goal: based on past data and current environment, forecast the arrival process of passengers to immigration control and their queueing, and find the *optimal* number of officers in charge
  $\{$*Note: these notions will be made more precise later*$\}$

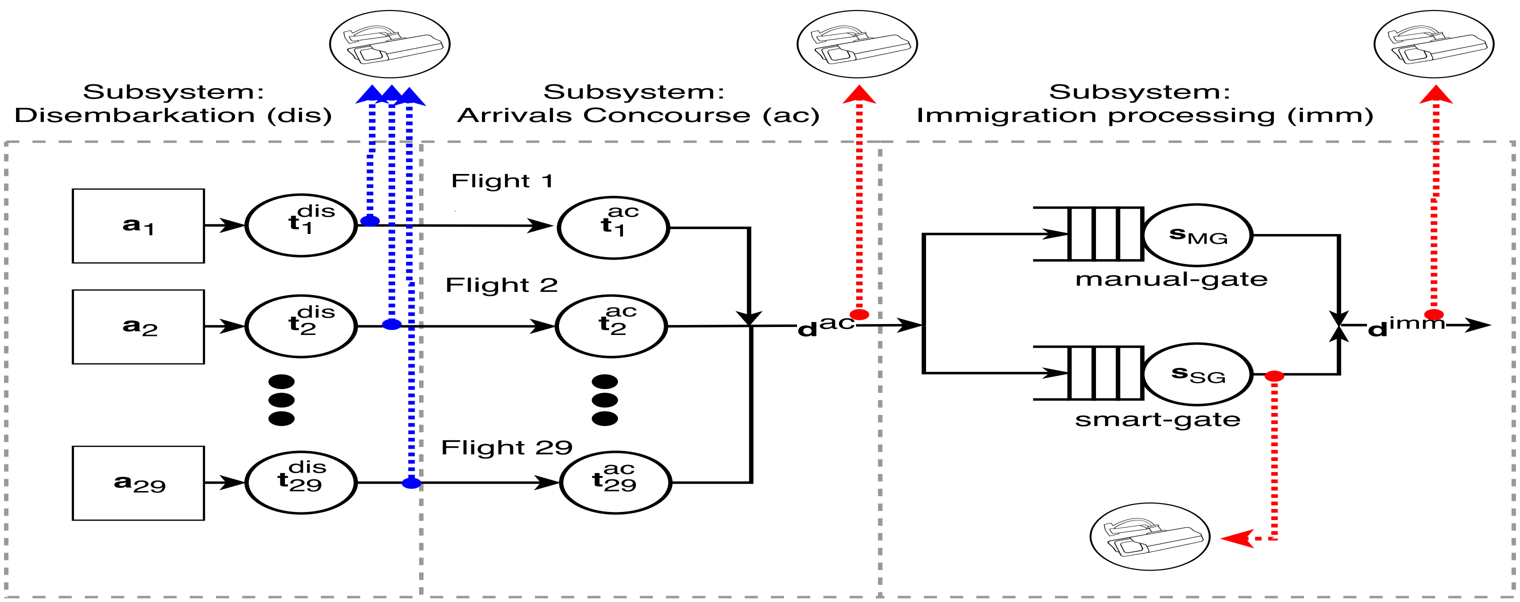
- Model complexity: although each step in the arrival process could be modelled stochastically, the lack of data for some of them makes the likelihood intractable

- Estimation complexity: in absence of a tractable likelihood, Approximate Bayesian Computation (ABC) is used to simulate data in the queueing process

- Simulation complexity: data simulation for this complex system is burdensome and queueing process has to be simulated efficiently $\Rightarrow$ R package `queuecomputer`

- Decision complexity: conflicting utilities (e.g., passengers' waiting time vs. officers' cost) $\Rightarrow$ future work

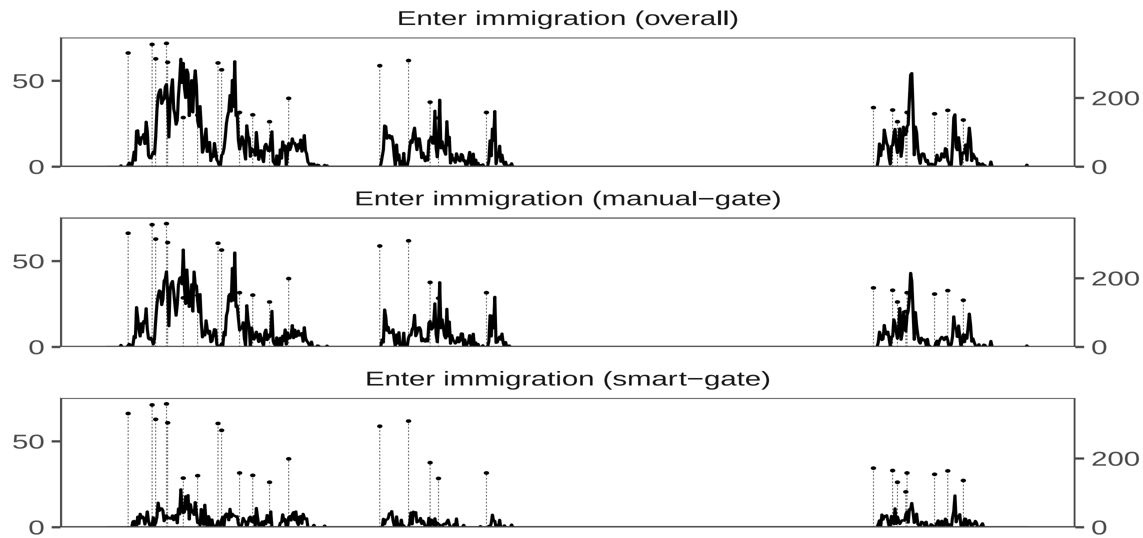
IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS



- Passenger arriving at gate a_i disembarks from the plane in a time t_i^{dis}, then he/she reaches the immigration area after a further time t_i^{ac} and here goes either to smart or manual gates, mostly depending on nationality, waiting for d^{ac} before being served and finally leaves the immigration time after an inspection lasted d^{imm}

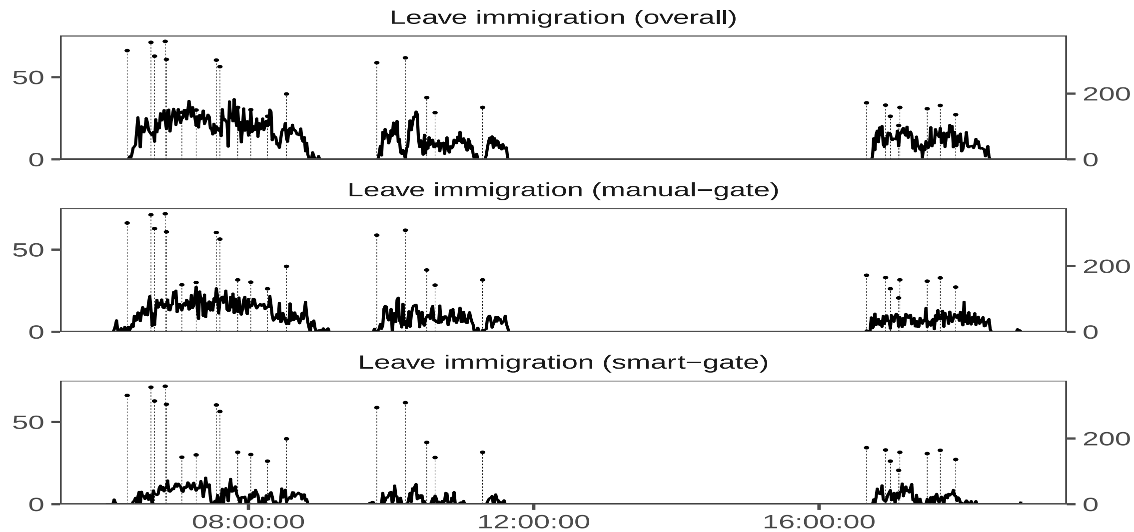
- CCTV cameras located at disembarkation, arrival and departure from the immigration area are counting the number of arriving passengers

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS



- X-axis: time of day
- Left Y-axis and black continuous lines: number of passengers entering at immigration
- Right Y-axis and dashed vertical lines: number of passengers arrived with a flight

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS



- X-axis: time of day
- Left Y-axis and black continuous lines: number of passengers leaving immigration
- Right Y-axis and dashed vertical lines: number of passengers arrived with a flight

QUEUES AND DYNAMIC QUEUEING NETWORKS

- Customer $j = 1, 2, \cdots$ enters a system at the *arrival time* a_j and requires a *service time* s_j with a server (e.g., customers in bank, patients in a hospital, passengers at immigration control)

- Sets of arrival and service times, ordered by customer, denoted as $\mathbf{a} = (a_1, a_2, \cdots)$ and $\mathbf{s} = (s_1, s_2, \cdots)$ respectively

- Typically a server can serve only one customer at a time $\Rightarrow$ unavailable or available if serving a customer or not

- Customers have to wait in the queue if all servers are unavailable $\Rightarrow$ *waiting times* denoted as $\mathbf{w} = (w_1, w_2, \cdots)$ (with $w_j = 0$ if the customer j arrives when a server is available)

- After being served, a customer leaves the system $\Rightarrow$ *departure times* denoted as $\mathbf{d} = (d_1, d_2, \cdots) \Rightarrow d_j = a_j + w_j + s_j$

QUEUES AND DYNAMIC QUEUEING NETWORKS

- Queueing system often summarized by six characteristics $A/S/c/K/M/R$, using Kendall's (1953) notation
  - A and S: forms of arrival and service processes, respectively
  - c number of servers
  - K finite or infinite capacity of the system
  - M finite or infinite customer population
  - R service discipline
- Simple examples
  - $M/M/1$
  - $GI/M/1$ and $GI/M/c$ Systems
  - $M/G/1$ System
  - $GI/G/1$ Systems

QUEUES AND DYNAMIC QUEUEING NETWORKS

Quantities of interest

- Arrival and service processes
- Client's perspective
  - Waiting time in queue
  - Size of the waiting queue
- Server's perspective
  - Busy period
  - Idle time between services
- $\Rightarrow$ Introduction of performance measures

QUEUES AND DYNAMIC QUEUEING NETWORKS

- Bayesian estimation possible for simple queueing systems, e.g.
  - $M/M/1$ (Armero and Bayarri)
  - $M/G/1$ (Rios Insua, FR and Wiper using mixtures of gamma distributions)
- Likelihood function unavailable or very difficult to derive $\Rightarrow$ likelihood-free methods may be required
  - Discrete Event Simulation (DES): interarrival and service times simulated over a sufficiently large time period
  - Approximate Bayesian Computation (ABC)
- Queueing network: customers transition between queueing systems
- Dynamic queueing network: queueing network with varying arrival rate
- Goal: develop a robust algorithm which will work with noisy measurements in queueing network with varying arrival rate

QUEUECOMPUTER

- R package *queuecomputer* developed by Ebert
  - algorithm denominated *queue departure computation* (QDC)
  - computationally efficient method for simulating departure times from a very general set of queueing networks
  - remarkable speedups of more than 2 orders of magnitude are observed relative to the popular DES packages *simmer* and *simpy*
  - output replicated from these packages to validate the package
  - key element in ABC applied to queueing networks

QUEUECOMPUTER

- Simulation for single server queueing system based on algorithm by Lindley (1952)
  - for i-th customer: generation of arrival time a_i and service time s_i
  - $\Rightarrow$ computation of departure time $d_i = \max(a_i, d_{i-1}) + s_i$
  - *customer either waits for a server or the server waits for a customer*
- Lindley's algorithm extended to multi-server systems by Krivulin (1994)
- QDC (queue departure computation) applies to very general multi-server systems, including
  - tandem queues (customers/tasks through ordered series of queues before departing the system)
  - parallel queues (customers/tasks partitioned into separate queueing systems)
  - fork/join queues (a task forked into a number of subtasks to be completed by distinct parallel servers and task can only depart the system once all subtasks have arrived at the join point, unlike the parallel queues)

QUEUECOMPUTER

- $\Rightarrow d_i = \max(a_i, d_{i-1}) + s_i$
  - $\mathbf{d} = (d_1, d_2, \cdots)$ grows with each new customer $\Rightarrow$ scalability problem (for $K > 1$ servers all departure times of past arrivals are needed to know when a server becomes available for the i-th customer)
  - computational complexity $O(n^2)$, for n customers
- Goal of `queuecomputer` and QDC: reduce the computational complexity
- First feature of `queuecomputer` and QDC: simulation of departure times for fixed number K of servers
- Algorithm able to simulate any queue of the form $G(t)/G(t)/K/\infty/n/FIFO$
  - general form for inter-arrival and service distributions, also varying over time
  - K servers for a population of n customers
  - unlimited capacity for waiting queue
  - customers served according to the ”First in - First out” policy

QUEUECOMPUTER

- Key idea of QDC: arriving customer choosing server which becomes available first
- i-th customer observes set of times $b_i = \{b_{ik} | k = 1, \ldots, K\}$ representing times when each server will be available next
- i-th customer selects the earliest available server $p_i = \text{argmin}(b_i)$ from b_i
- Departure time for i-th customer: $d_i = \max(a_i, b_{p_i}) + s_i$
  (server must wait for customer or vice versa)
- Each b_{ik} could be given either by the remaining service time of a customer in the server or by the sum of such time and the service times of the customers who queued at such server based on their b_{p_i}
- QDC algorithm for a fixed number of servers pre-sorts the arrival times and considers $\mathbf{b}$ as a continually updated K length vector representing the state of the system, instead of assigning b_i's to each customer i to form a matrix $\mathbf{b}$ of size $K\times$ number of costumers

QUEUECOMPUTER

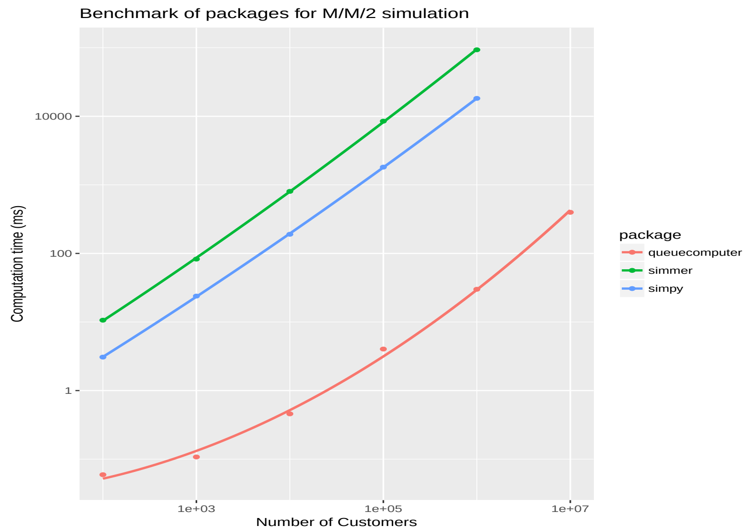
- Number of servers available to customers could change over time
- Realistic situations where more servers are scheduled for busier times of the day
- Second feature of `queuecomputer` and QDC: simulation of departure times for varying number K of servers
- $K(t)$: number of servers at time t
- Number of open (available) servers throughout the day as a step function
- Time partitioned by L knot locations $\mathbf{x} = (x_1, \cdots, x_L) \in \mathbb{R}_+^L$ into $L + 1$ epochs $(0, x_1], (x_1, x_2], \cdots, (x_L, \infty)$
- Number of open servers in each epoch represented by a $L + 1$ length vector $\mathbf{y} = (y_1, \cdots, y_{L+1}) \in \mathbb{N}_0^{L+1}$
- Step function in input and determined by the user, changeable before the simulation but not during it, like arrival and service times $(\mathbf{a}, \mathbf{s})$

QUEUECOMPUTER

- Server k can be closed setting $b_k = \infty$ ensuring that no customer can use it
- Server k can be opened at time t setting $b_k = t$ allowing customers to use it
- No limitation about generation of knot locations $\mathbf{x}$ and number $\mathbf{y}$ of open servers in each epoch
- Algorithm available in the paper
- Algorithm based on number of open servers at each epoch but unable to select which servers are to be closed or opened
- Less efficient algorithm developed to allow for the selection of the servers to open or close
- `queuecomputer` available at

  https://cran.r-project.org/web/packages/queuecomputer/index.html

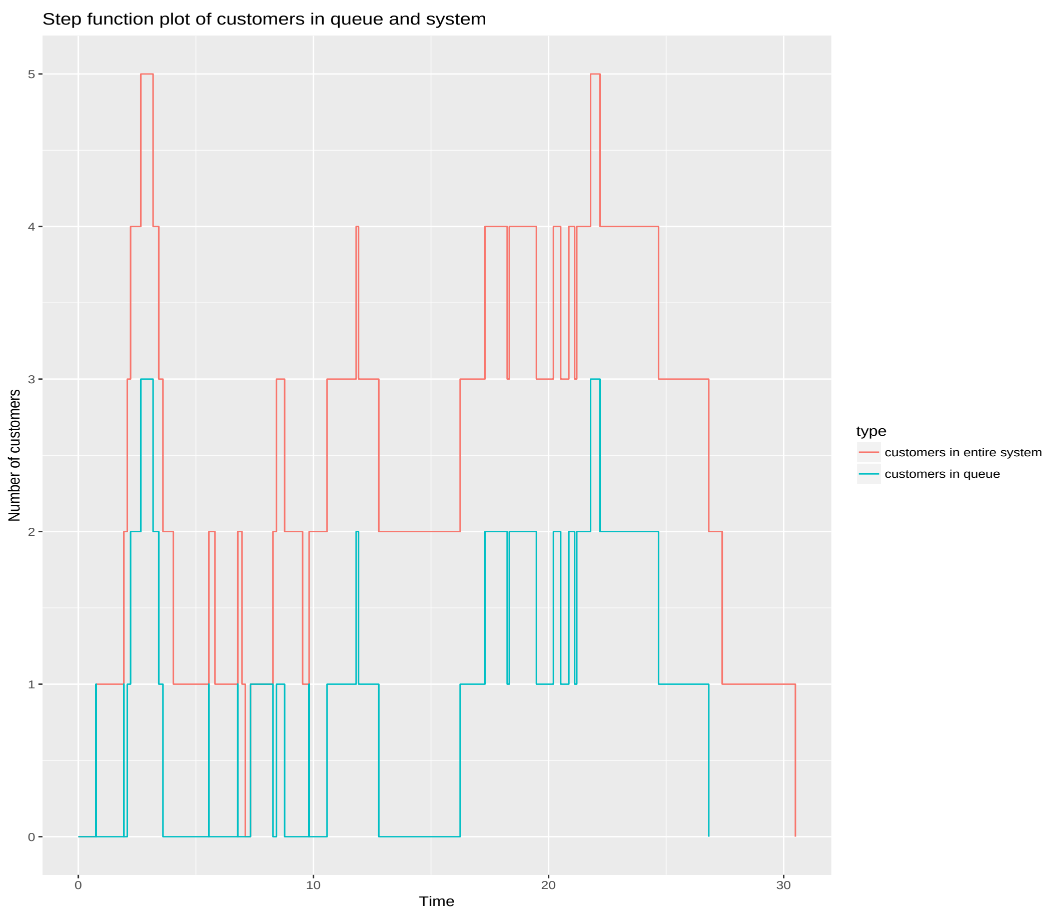
QUEUECOMPUTER



- $M/M/2$ queue: Computation time in milliseconds for varying numbers of passengers for each DES/queueing package
- Each package returns exactly the same set of departure times, since the same arrival and service times are supplied
- Computation time reported here is the median time of 100 runs for each number of customers and each package. Intel (R) Core(TM) i7-6700 CPU @ 3.40GHz running Debian GNU/Linux

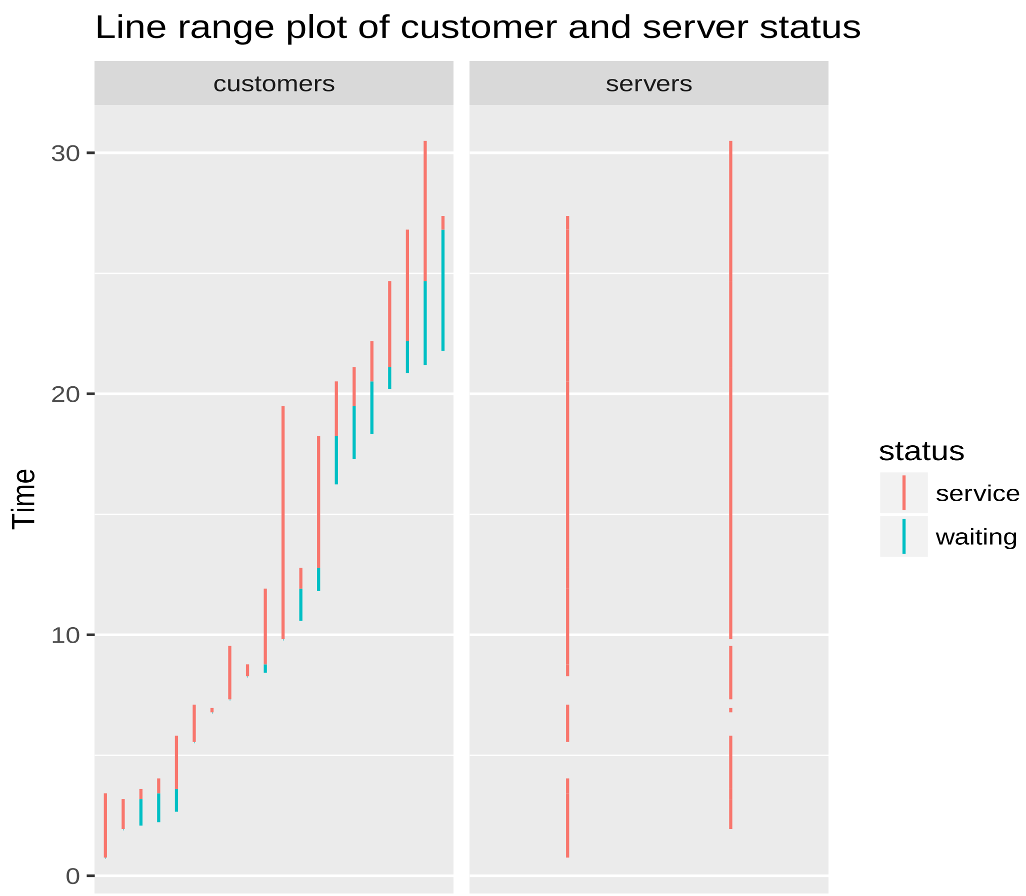
QUEUECOMPUTER: SIMULATIONS

Queue length and number of customers in system over time ($K = 2$ servers)



QUEUECOMPUTER: SIMULATIONS

Waiting and service times for each customer ($K = 2$ servers)



APPROXIMATE BAYESIAN COMPUTATION (ABC)

- Interest in the posterior distribution of parameter θ given observations $\mathbf{y}$
- Markov chain Monte Carlo (MCMC) relies on evaluation of the likelihood $f(\mathbf{y}|\theta)$
- Evaluation of $f(\mathbf{y}|\theta)$ too costly for some complex statistical models
- $\Rightarrow$ likelihood-free computational methodologies such as ABC
- ABC sampler $\Rightarrow$ parameter proposals θ^* to generate $\mathbf{x} \sim f(\mathbf{x}|\theta^*)$ to compare with observed data $\mathbf{y}$ via a distance function ρ
- Common practice: define ρ as a distance between lower-dimensional summary statistics S and accept θ^* if $\rho < \epsilon$
- Accepted $\theta^* \Rightarrow$ draws from the ABC posterior $\pi_{\text{ABC}}(\theta|\mathbf{y})$, approaching true posterior $\pi(\theta|\mathbf{y})$ for sufficient S and $\epsilon \downarrow 0$
- Critical choice of sufficient S: low dimension $\Rightarrow$ reduced computational time but possible loss of information

APPROXIMATE BAYESIAN COMPUTATION (ABC)

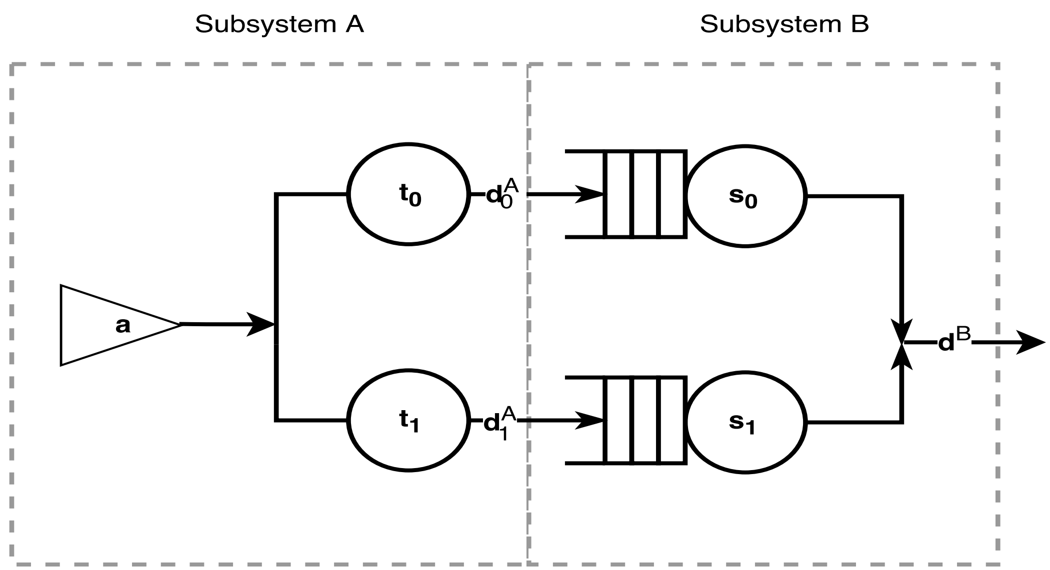
- Instead of a sufficient statistics we consider maximum mean discrepancy (MMD) between the two samples
- Given two r.v.'s X and Y and a class $\mathcal{F}$ of functions, then MMD is defined as $\sup_{f \in \mathcal{F}} (\mathbf{E}_X[f(X)] - \mathbf{E}_Y[f(Y)])$
- Gretton et al. (2012) consider $\mathcal{F}$ as the unit ball in a Reproducing Kernel Hilbert Space
- Biased empirical estimate of MMD for two samples $\mathbf{x}$ and $\mathbf{y}$ given by

$$\widehat{\rho}_{\text{MMD}}(\mathbf{x}, \mathbf{y}) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j)$$

  - m length of $\mathbf{x}$
  - n length of $\mathbf{y}$
  - k kernel function, here Gaussian: $k(x, y) = \exp\left(-\frac{(x-y)^2}{2\sigma_k^2}\right)$, with σ_k fixed tuning parameter

SIMULATION: ILLUSTRATIVE EXAMPLE

System with parallel queues



- $\mathbf{a}$: arrival times
- $\mathbf{t}_i$, $\mathbf{d}_i^A$ and $\mathbf{s}_i$: transition, arrival and service times to queue i, $i = 0, 1$
- $\mathbf{d}^B$: departure times

SIMULATION: ILLUSTRATIVE EXAMPLE

System with parallel queues

- Arrival times $\mathbf{a}$ drawn from known, dynamically varying density function $f_{\mathbf{a}}$
- Upon arrival, i-th customer routed to one of the queueing systems (0 or 1) with Bernoulli r.v. r_i s.t. $P(r_i = 0) = p$
- Transition time to the queueing system after routing assignment: $t_i \sim \text{Gamma}(\alpha, \beta)$
- $\Rightarrow$ i-th customer arriving to the queue at time $d_i^A = a_i + t_i$
- Exponential service s_i with parameter λ_0 or λ_1 depending on the i-th customer's route
- Number of servers in the two routes: K_0 and K_1
- Departure times d_i^B's from each queueing system computed deterministically using QDC

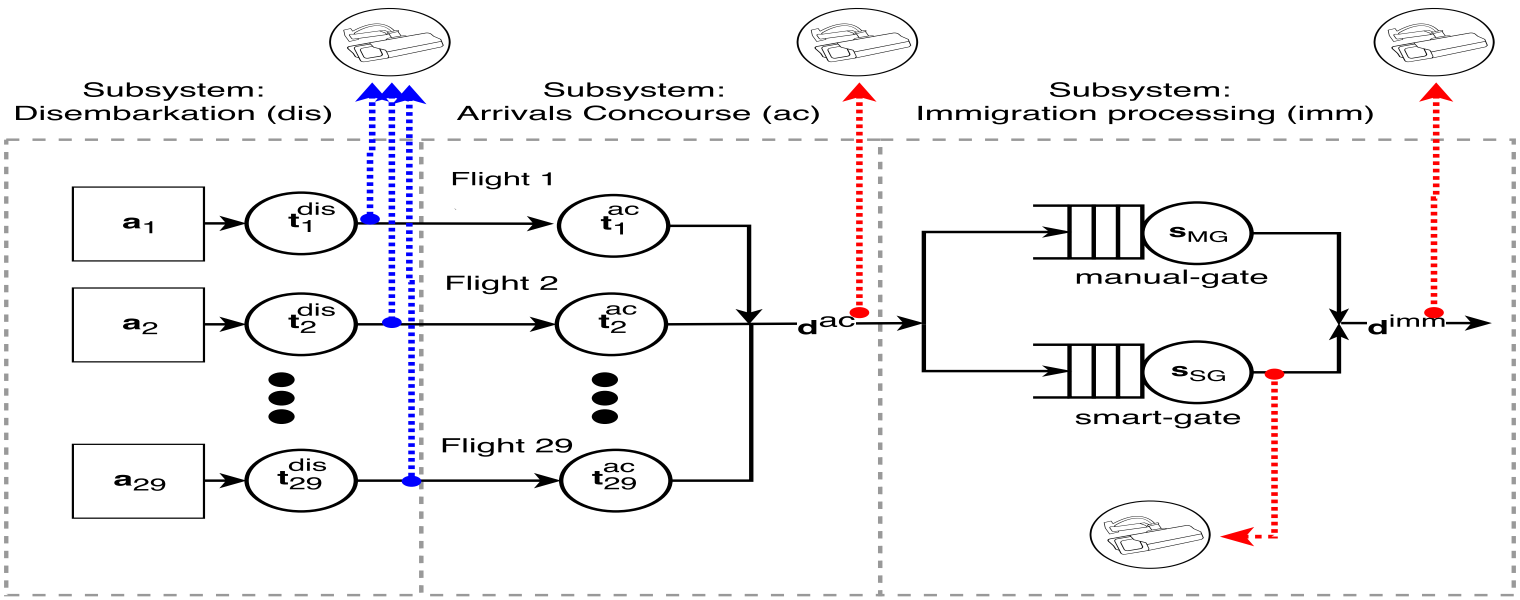
SIMULATION: ILLUSTRATIVE EXAMPLE

Two possible uses of QDC:

- Simulations of times for fixed parameters $\Rightarrow$ computations of performance measures, like length of queues, waiting time, number of customers in the system, idle periods

- Estimation of $\theta = (p, \alpha, \beta, \lambda_0, \lambda_1)$ when some d_i^A's and d_i^B's, i.e. $\mathbf{y}$, are available
  - likelihood function $f(\mathbf{y}|\theta; f_\mathbf{a}, \mathbf{K})$ cannot be evaluated
  - $\Rightarrow$ θ estimated embedding a queueing simulation within an ABC sampler
    * values $\tilde{\theta}$ generated from prior
    * QDC used to generate $\mathbf{x}$, conditional on $\tilde{\theta}$ and known inputs $f_\mathbf{a}$ and $\mathbf{K}$
    * MMD computed between $\mathbf{x}$ and $\mathbf{y}$ and $\tilde{\theta}$ accepted or rejected of
    * $\Rightarrow$ approximate posterior distribution on θ

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS



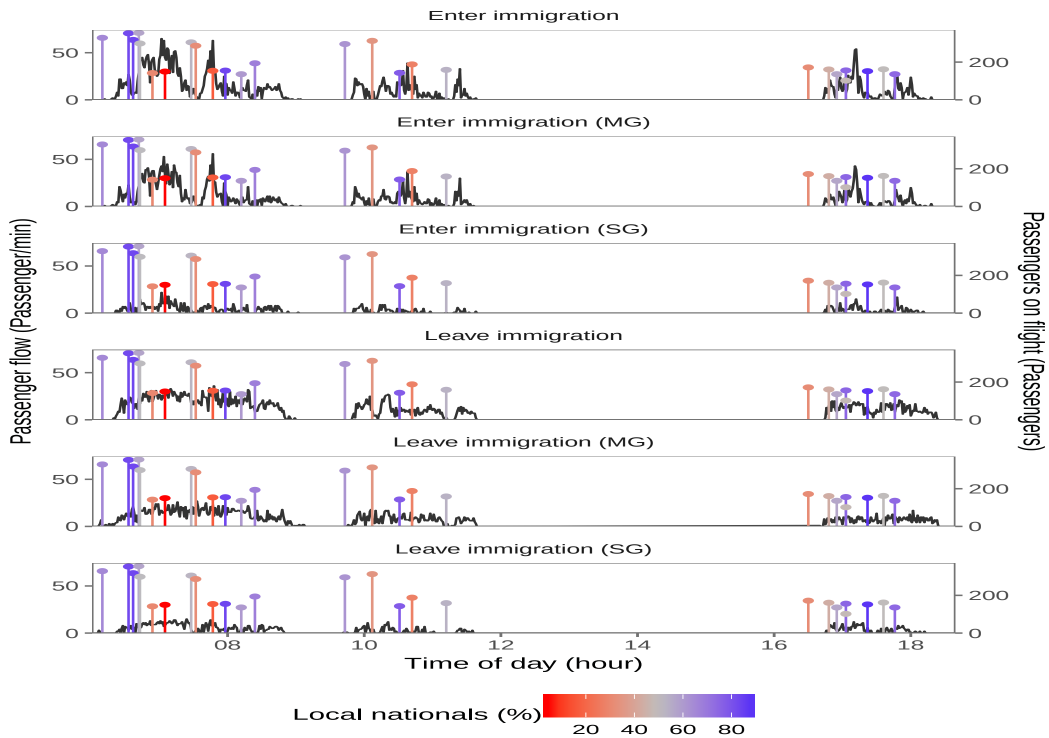
- Passenger arriving at gate a_i disembarks from the plane in a time t_i^{dis}, then he/she reaches the immigration area after a further time t_i^{ac} and here goes either to smart or manual gates, mostly depending on nationality, waiting for d^{ac} before being served and finally leaves the immigration time after an inspection lasted d^{imm}

- CCTV cameras located at disembarkation, arrival and departure from the immigration area are counting the number of arriving passengers

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

- Data based on 29 flights and 5453 passengers in total
- Information on each flight i
  - arrival time: a_i
  - distance from arrival gate to immigration control: m_i
  - number of passengers on the flight: j_i
  - % of passengers who are local nationals, as opposed to foreign nationals: p_i^{nat}
- Information on resource levels assigned to each queueing system
  - number of machines at smart gates (SG): $\mathbf{K}_{\text{SG}}$
  - number of staff members at manual gates (MG): $\mathbf{K}_{\text{MG}}$

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS



- Black lines: Measured passenger flows for the arrivals terminal
- Coloured vertical lines: Number of passengers on each flight arranged by flight arrival time, with colour representing the proportion of local nationals

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

- Gamma distribution for disembarkation times
- Gamma distribution for time to immigration control
- Bernoulli distribution for being national or foreigners
- Bernoulli distribution for selecting smart or manual gates, depending on nationality
- Exponential distribution for service time depending on the selected gate
- Vaguely informative priors on parameters (uniform, in general)
- Departure times obtained via QDC
- Data provided by CCTV footage, unable to follow individuals moving in the terminal and distinguish passengers from different flights

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

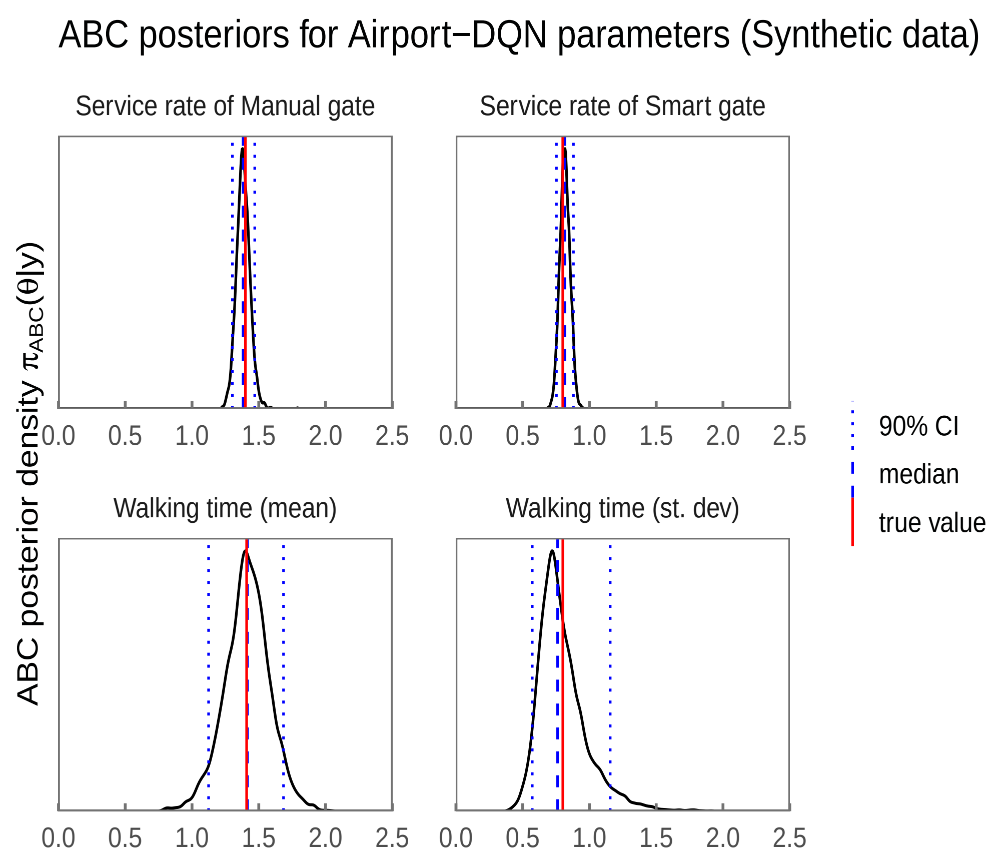
We did three analyses:

- Synthetic (simulated) data to check the validity of the model

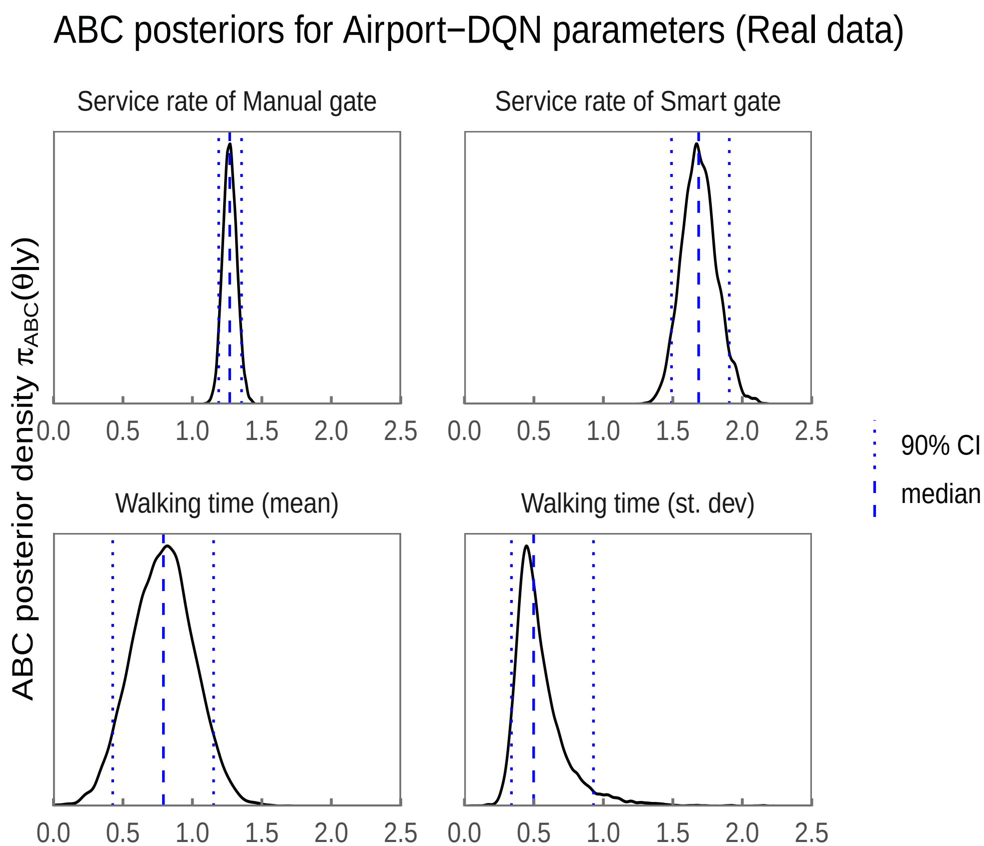
- Real data (not publicly available)

- Realistic (perturbed) data (journal's policy about reproducibility of research)

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

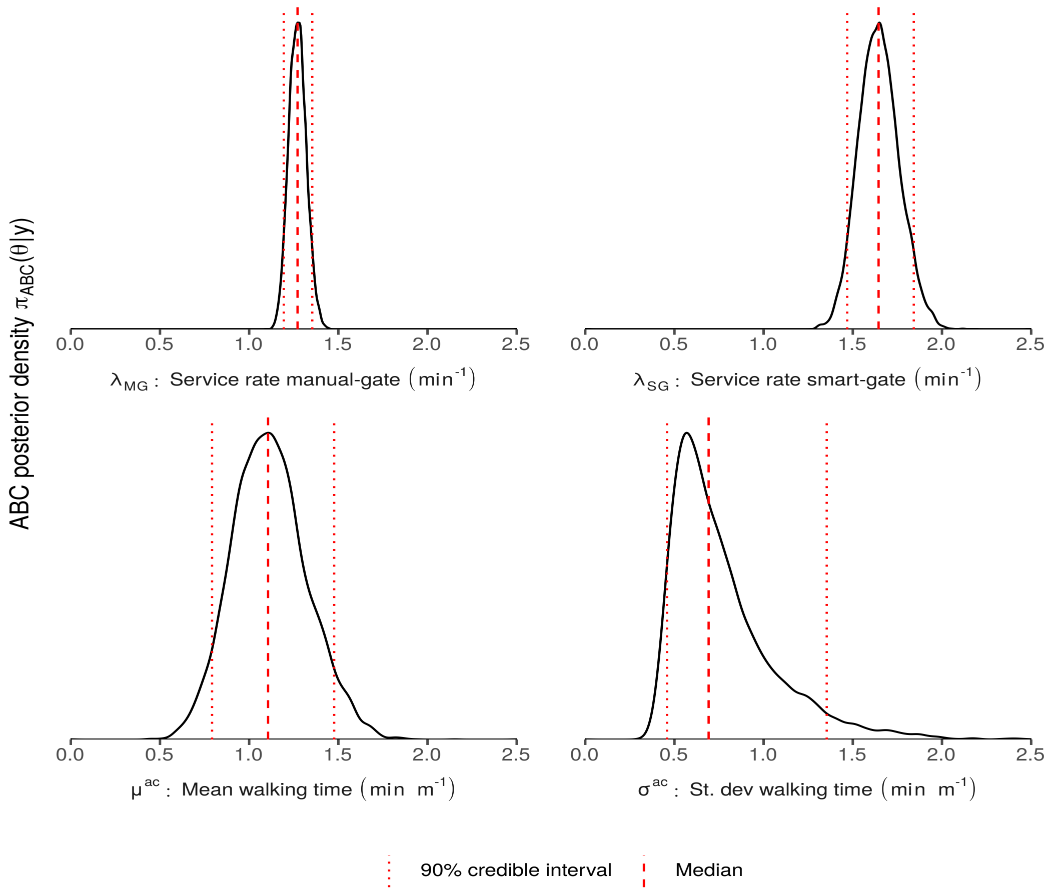


IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

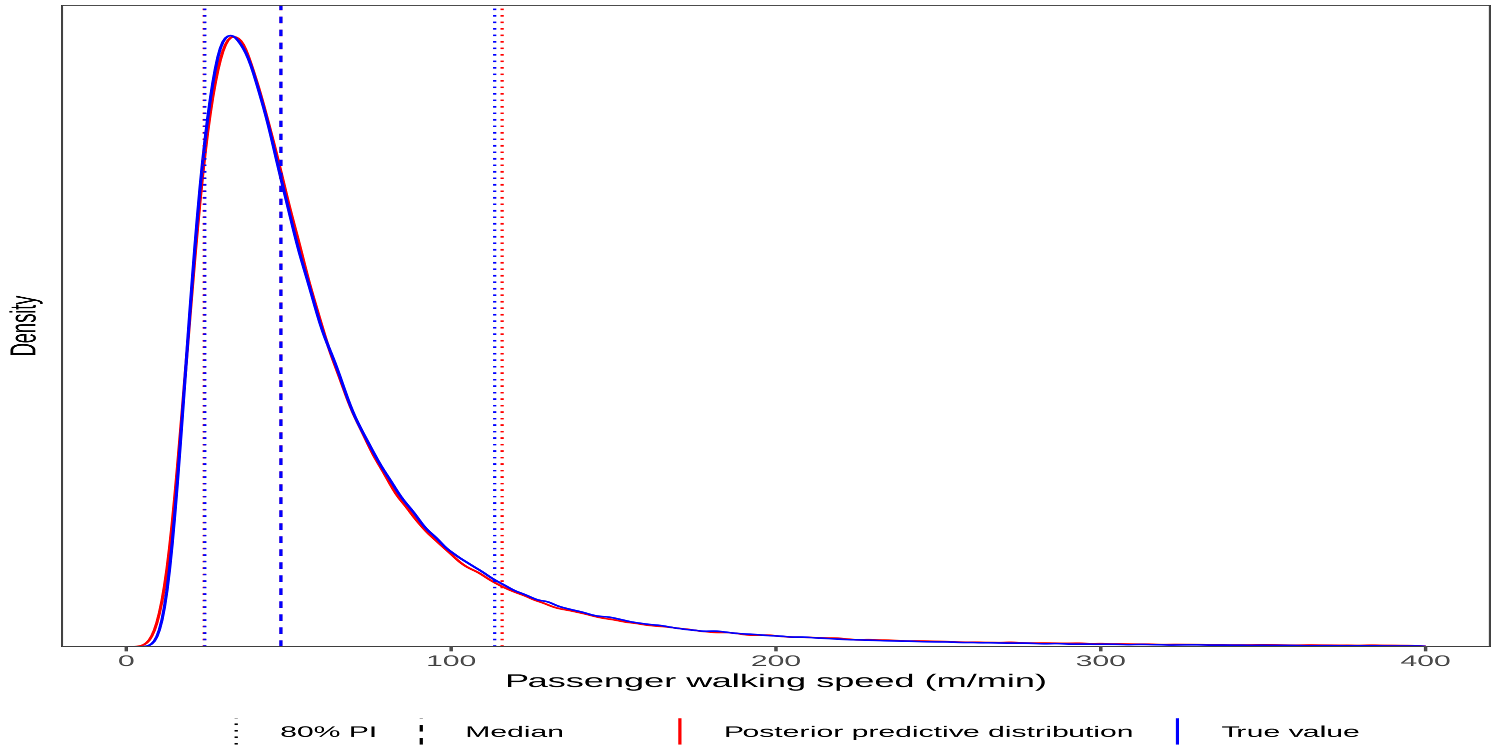


IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

ABC posteriors for Airport - DQN parameters (Perturbed data)



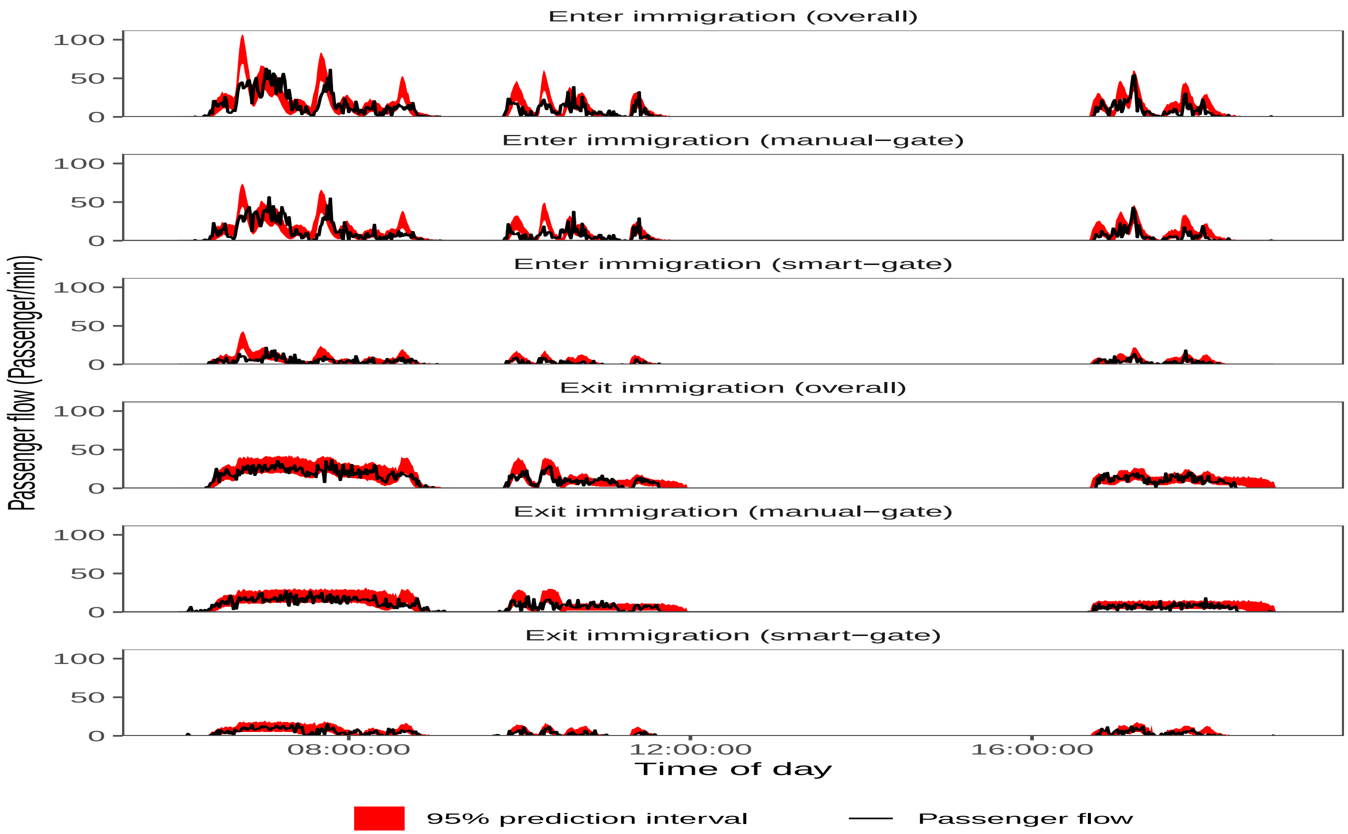
IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS



- Posterior predictive distribution of walking speeds (synthetic data)
- Findings for real data in agreement with literature

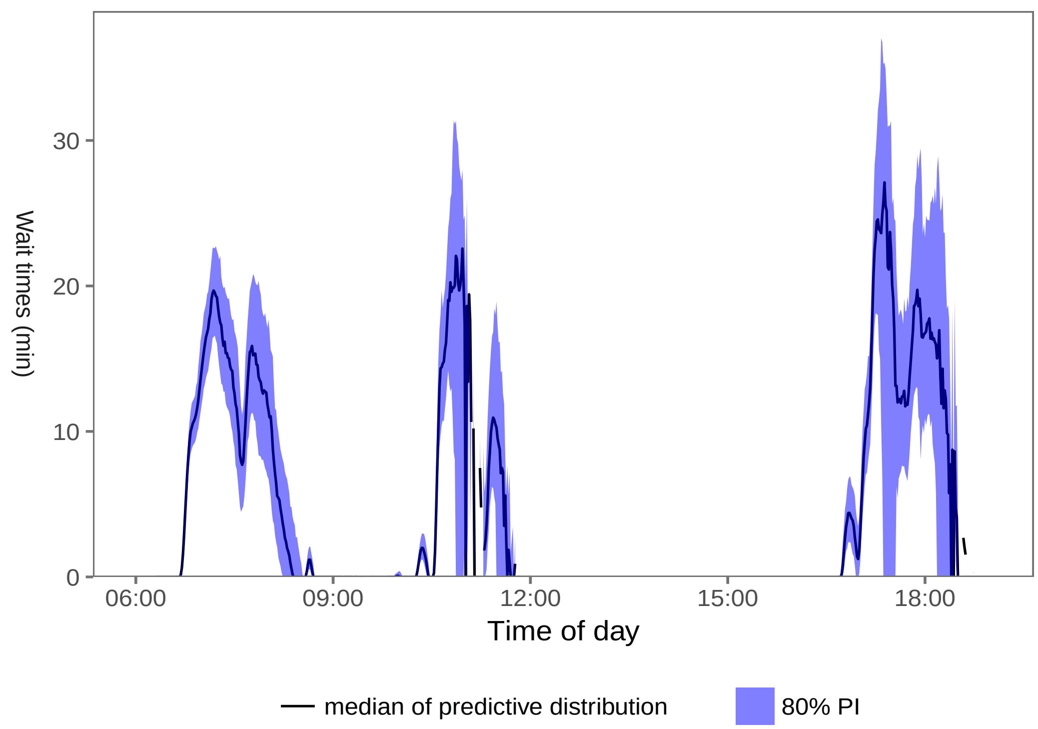
IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

Comparison between passenger count data $\mathbf{y}$ and predictive intervals from model realisations $\mathbf{x}$ based on the posterior distribution to assess validity of the model



IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

Performance measures (here waiting times for a new passenger) with predictive intervals using draws from $\pi(\theta|\mathbf{y})$ along with flight and resource schedule



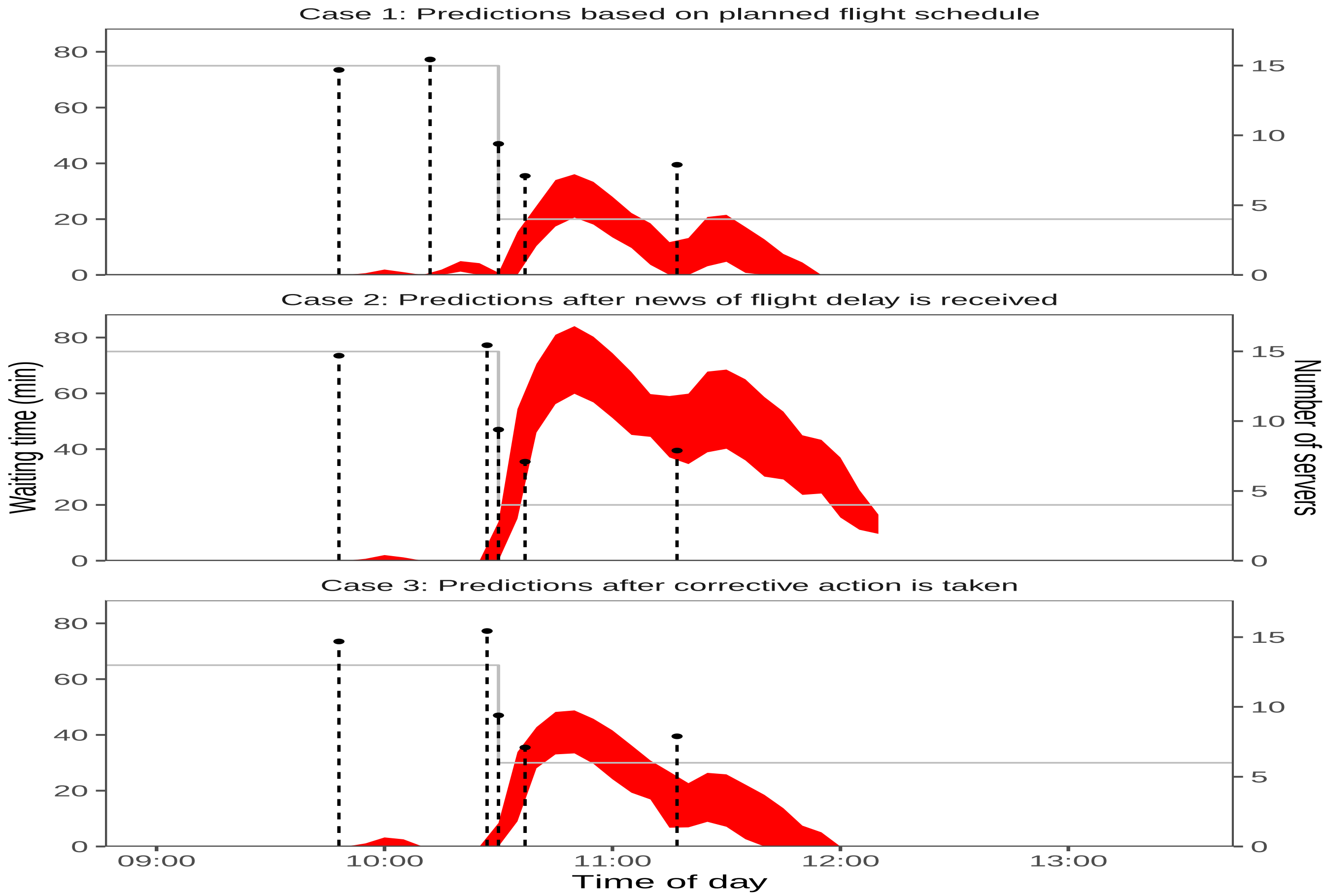
IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

Prediction of future performance measures as a decision support tool

- Case 1 (normal): predictions follow from planned flight schedule
- Case 2 (disruption): second flight delayed by 15 minutes $\Rightarrow$ increase of median waiting time at 10.50 from 28 to 71 minutes and median queue length from 146 to 363 passengers
- Case 3 (intervention): two staff moved from earlier to later shift $\Rightarrow$ decrease of median waiting time to 41 minutes and median queue length to 319 passengers

IMMIGRATION QUEUE AT INTERNATIONAL ARRIVALS

Prediction intervals (95%) for waiting times at the manual gates



CRITICAL ASPECTS

- Although our estimated walking speed (91.9 m/min) is very close to the one obtained by ad hoc studies (80.5 m/min), we have not considered factors like retail outlets, bathroom facilities, family groups and congested passenger flows

- When comparing passenger counts in the real and ABC simulated data, many but not all the peaks overlap with the prediction interval: a small translation in peak positions can have a large effect on a functional distance estimator

- We chose MMD to assess distance between observed and simulated data whereas we found unsatisfactory results when using the Wasserstein distance
  $$W(\mu, \nu) = \sup \left\{ \left| \int f d\mu - \int f d\nu \right| : f \text{is Lipschitz} \right\}$$
  Is there an *optimal* distance for ABC in dynamic queueing systems?

- Work limited to queues at immigration control, but other queues arise at security control and check-in, just to mention two

- Best estimation strategy in tandem queues: all the parameters at once or one at the time (e.g. first the walking rate to the immigration control and then the service rate)?

FUTURE AND POSSIBLE WORK

- Optimal allocation of resources (staff, smart gates)
- Multicriteria Decision Analysis: conflicting interests
- Adversarial Risk Analysis: game between resources planner and incoming passengers
- Sensitivity w.r.t. variations in priors and simulated data
- Selection of optimal distance for ABC in dynamic queueing systems
- Work limited to queues at immigration control so far, but other queues arise at security control and check-in, just to mention two
- Registration of curves (e.g. through shift of peaks) to improve efficacy of ABC

BIBLIOGRAPHY

- Ebert, A., Wu, P., Mengersen, K. and Ruggeri, F. (2020), Computationally Efficient Simulation of Queues: The R Package queuecomputer. *Journal of Statistical Software*, 95, 1-29.

- Ebert, A., Dutta, R., Wu, P., Mengersen, K., Ruggeri, F. and Mira, A. (2019), Likelihood-Free Parameter Estimation for Dynamic Queueing Networks. Under revision

- Farr, A.C., Ruggeri, F. and Mengersen, K.L. (2018), Prior and Posterior Linear Pooling for combining expert opinions: uses and impact on Bayesian Networks. *Entropy*, 20(3):209.

- Farr, A.C., Simpson, D.P., Ruggeri, F. and Mengersen, K.L. (2020), Combining opinions for use in Bayesian Networks: a Measurement Error Approach. *International Statistical Review*, 88, 335-353.