MIT Open Access Articles

Small-area population forecasting in a segregated city using density-functional fluctuation theory

Massachusetts Institute of Technology

Small-area population forecasting in a segregated city using density-functional fluctuation theory

Yuchao Chen[1] · Yunus A. Kinkhabwala[2] · Boris Barron[1,3] · Matthew Hall[3,4] · Tomás A. Arias[1] · Itai Cohen[1]

Abstract

Policy decisions concerning housing, transportation, and resource allocation would all benefit from accurate small-area population forecasts. However, despite the success of regional-scale migration models, developing neighborhood-scale forecasts remains a challenge due to the complex nature of residential choice. Here, we introduce an innovative approach to this challenge by extending density-functional fluctuation theory (DFFT), a proven approach for modeling group spatial behavior in biological systems, to predict small-area population shifts over time. The DFFT method uses observed fluctuations in small-area populations to disentangle and extract effective social and spatial drivers of segregation, and then uses this information to forecast intra-regional migration. To demonstrate the efficacy of our approach in a controlled setting, we consider a simulated city constructed from a Schelling-type model. Our findings indicate that even without direct access to the underlying agent preferences, DFFT accurately predicts how broader demographic changes at the city scale percolate to small-area populations. In particular, our results demonstrate the ability of DFFT to incorporate the impacts of segregation into small-area population forecasting using interactions inferred solely from steady-state population count data.

Keywords Small-area forecasts · Residential choice · Migration · Schelling model · Density-functional fluctuation theory · Segregation

Introduction

Forecasting neighborhood-scale population changes in response to residential migrations remains a challenge, with the potential to inform and significantly affect local planning of social and economic developments [46]. For example,

---

Yuchao Chen and Yunus A. Kinkhabwala have contributed equally to this work.

Extended author information available on the last page of the article

Springer

such forecasts could be used to achieve optimal allocation of educational, health and safety resources by determining the need for new schools [49], hospitals [22] and fire stations [35] in each neighborhood. In addition, such forecasts would be important for estimating housing demands [27], and might help combat socioeconomic inequalities by predicting the need for low-income housing developments [1] and public transportation [56, 57]. Despite this wide range of important potential applications, methods for accurate neighborhood-scale population forecasts remain limited.

At regional scales, there are already numerous methods for predicting population change and dynamics. Traditionally, the demographic equation is applied to different cohorts [24, 26, 37, 44, 46, 47, 52]. In particular, estimates of birth and death rates from past data are already quite accurate. Models with different amounts of sophistication have also been developed to estimate the migration rates from available data [43]. For example, the gravity model [14, 18, 23, 25, 36, 38] and the gravity-like Weidlich-Haag Migratory Model [19, 53–55] fit migration data by including relative preferences of origin and destination regions as well as preferences to make moves to closer locations. The challenge, then, is how to relate forecasts at the regional level to forecasts at the neighborhood scale.

One of the many hurdles for making such relations is that, at the neighborhood scale, drivers of segregation can significantly affect the resulting distributions. At regional scales, most methods either ignore the drivers of segregation [38] or assume some simple forms for such segregation effects [19]. These simplifications are most likely justified for regional or national population forecasts, because the drivers of segregation, such as economic status, social preferences and housing policy [10, 16], are likely to average over large scales. Such simplifications, however, could lead to substantial errors if we apply them to make small-area forecasts.

While many small-area population forecast methods have been developed and tested, none to date explicitly account for the effects of residential segregation. [59] recently collected a thorough review of these small-area methods, including extrapolative, cohort-component, and small-area microsimulations. Some methods referenced here account for aspects that might be of special importance to small-area dynamics such as land use, roads, urban accessibility [29], water body, country borders [4], demographic and socioeconomic characteristics [8]. However, residential segregation remains a key feature of American society and a potent driver of residential mobility at the neighborhood scale. For example, residential tipping, the process in which in-migrations or out-migrations of neighborhoods are driven by the current racial compositions, can lead to dramatic differences in neighborhood population changes [7]. Thus, there is a need to explore methods that can account for segregation effects in order to provide accurate forecasts of neighborhood migration.

Although there is a rich history of methods aimed at quantifying segregation and understanding its causes, it is unclear how to use most of these methods to create forecasts of small-area populations. One general approach is to quantify segregation based on numerical indices, so-called segregation indices, to characterize the degree of segregation in a city [12, 15, 31–34, 39–41, 58]. Such indices have been essential to understanding how segregation correlates with residential outcomes as well as potential drivers of segregation. Recently, [13] even used such indices to

extrapolate which neighborhoods are more likely to change their degree of segregation but stopped short of forecasting population changes, presumably because such indices are too coarse-grained to make accurate predictions. Another approach utilizes agent-based models [5, 6, 11, 17, 42, 45, 48, 51, 60–63], such as the well-studied Schelling model [5, 17, 45, 60–62], to determine the degree to which different proposed interactions lead to segregation and to investigate their dynamics. Such studies have shown that even slight preferences towards segregated neighborhood compositions can lead to drastic city-wide segregation and dynamic phenomena such as residential tipping. Since these models require *a priori* knowledge of the decision rules for migration that are challenging to determine in practice, they have found limited use for predicting quantitative trends in human populations [3]. Thus, despite great progress in understanding the nature of segregation, neither segregation indices nor agent-based models have led to widely adopted methods for predicting population dynamics at the small-area scale.

Recently, a new statistical physics method called Density-Functional Fluctuation Theory (DFFT) was developed to make steady-state predictions of how crowds will distribute in different environments [30]. In short, DFFT is a top-down data-driven approach that extracts functions to separately quantify effective social and spatial preferences from observed fluctuations in small-area population densities. By combining these functions, DFFT is able to predict population distributions in new environments, suggesting it could be a very promising approach to generate small-area forecasts. More recent work by Barron et al [2] connected DFFT to traditional measures of segregation, demonstrating that a large variety of segregation indices can be directly computed from functions arising in DFFT. This suggests that DFFT is uniquely positioned to relate measures of segregation with population forecasting, bridging these traditionally distinct branches of demographic research.

This work represents the first application of DFFT beyond steady-state considerations by extending DFFT to predict small-area data *over time* for multi-group systems (*i.e.,* multiple types of agents in a system evolving in time far from equilibrium). In particular, we endeavor to forecast the number of individuals in small-area units of a simulated city that is driven by a Schelling-type model. While this city is not expected to capture complexities in real-world segregation, it—crucially—allows us to demonstrate the efficacy of DFFT for small-area forecasting in a controlled setting.

Notably, DFFT relies *only* on small-area population counts and does not require migration data. While population count data is widely accessible (e.g., Census data), migration data is often private. Census data is also accompanied by additional methodological challenges including non-uniform populations across small-areas, changes in small-area geometries over time, and time-varying changes in behavior. These challenges require methodologies that are beyond the scope of this paper. As such, we rely here solely on simulated data to test the ability of DFFT to capture interactions from small-area population counts and to then use this information to make forecasts under new conditions.

In this work, we first consider a Schelling-type model to simulate a city where agents exhibit both social and spatial drivers of segregation. Whereas the social drivers are directly dependent on the neighborhood composition, spatial

drivers are neighborhood-specific and could represent neighborhood housing costs, crime, school quality, or other perceived neighborhood characteristics [9, 50]. We then use these simulations to create steady-state population data (i.e., when small-area distributions no longer change drastically over time) for our analysis (Fig. 1a; Sect. 2). Next, we apply DFFT to this data and extract functions describing the effective spatial and social preferences of the population (Fig. 1b; Sect. 3) while using no knowledge of the underlying agent behaviors other than their observable effects on small-area agent distributions. Notably, this quantification of preferences is generic so that it can in theory capture cumulative effects of the drivers of segregation without making specific assumptions about their properties or relative strengths.

We proceed by introducing a sudden regional-scale population change, resulting in segregation-driven small-area population changes over time. (This is reminiscent, for example, of how following widespread damage from Hurricane Katrina, the sudden population influx to Houston, Texas, resulted in segregation-driven intra-city migrations, e.g. 'White flight', over the next several years [21].) We consider a drastic population change for this demonstration because the resulting migrations could present greater discrepancies with predictions than those arising from small population changes, thereby providing a stress test for our predictive framework.

To *predict* the time evolution of small-area population data following the sudden regional-scale population change, we develop a time-dependent version of DFFT (TD-DFFT). TD-DFFT, while relying on functions extracted from the original steady-state data, extends the scope of DFFT by enabling predictions even in scenarios where the system is outside the steady state. We then compare the predictions from TD-DFFT to the data generated by the underlying Schelling-type model (Fig. 1c; Sect. 4). Finally, we predict the new steady-state joint densities resulting from the sudden regional-scale population change, either through numerical computation using our novel TD-DFFT or analytic calculation using the DFFT functions. These predictions are then compared to the observed data for the new steady state (Fig. 1d; Sect. 5).

In the following sections we will describe the Schelling model and DFFT framework in detail. The variables will be defined as they are introduced, but for ease of reference Table 1 summarizes the notation used throughout this paper.

Modified schelling simulation

To generate sample population data, we use a dynamic Schelling-type agent-based model [17] modified to include spatial dependence. In this model, two types of agents, 1000 red and 1000 blue, make probabilistic moves to new empty cells on a 60-by-60 lattice grid with periodic boundary conditions (Fig. 2a). The moves are based on changes in utility functions that specify social ("Social Utility" U_R^{so}, U_B^{so}) and spatial ("Spatial Utility" U_R^{sp}, U_B^{sp}) preferences. In particular, at each step in time, we randomly choose an agent and an empty cell, and then the agent accepts the move to the empty cell with probability
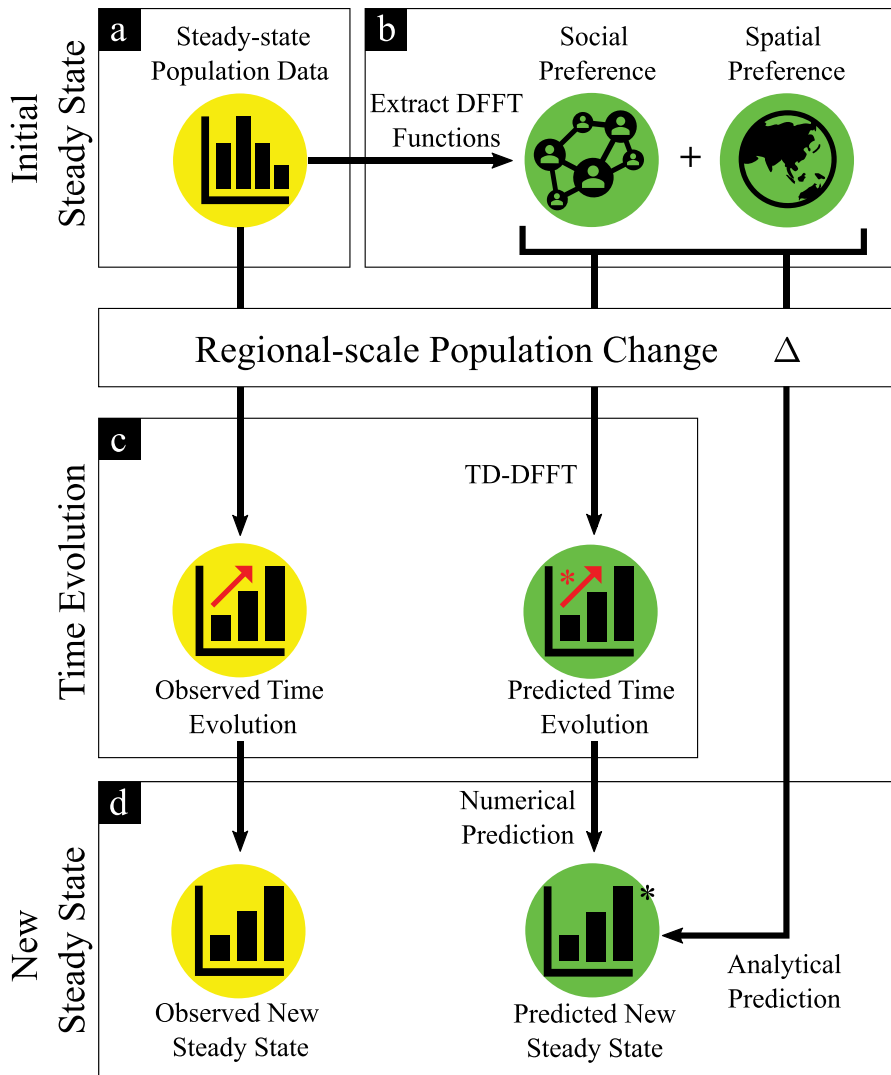
Fig. 1 General Workflow of Applying DFFT to Data of Segregated Populations. a Collect small-area steady-state population data in the form of probability distributions of local densities. Steady state is reached when the population distribution no longer changes drastically over time. In our example, we simulate the steady-state data from a Schelling model (yellow bubble). **b** Extract DFFT functions from steady-state data. The DFFT functions characterize social and spatial preferences separately. **c** After a population change, we predict the time evolution of small-area population data with time-dependent DFFT (TD-DFFT) using the extracted DFFT functions. We compare our prediction with the observed time evolution from the Schelling model simulations (yellow bubble). **d** We predict the new small-area steady state after the regional-scale change either numerically using TD-DFFT or analytically using DFFT functions. We compare our prediction with the observed new steady state of the Schelling model simulations (yellow bubble)

Table 1 Definition of variables

Variable	Meaning	Usage
$P_{\text{Schelling}}$	Probability of moving for each Schelling step	Schelling
U_R^{so}	Social Utility for a red agent	
U_B^{so}	Social Utility for a blue agent	
U_R^{sp}	Spatial Utility for a red agent	
U_B^{sp}	Spatial Utility for a blue agent	
x	Position in Schelling city	
N_R^{ne}	Number of red agents in the 8-connected neighborhood	
N_B^{ne}	Number of blue agents in the 8-connected neighborhood	
b	Block index	Schelling & DFFT
b_{tot}	Total number of blocks in a city	
t	Time	
s	Number of cells (spaces) in a block	
s_{tot}	Total number of cells (spaces) in a city	
N_{tot}	Total number of agents in a city	
N_R^{tot}	Total number of red agents in a city	
N_B^{tot}	Total number of blue agents in a city	
P_b	Probability Distribution of agents in block b	
$N_{R,b}$	Number (Density) of red agents in block b	
N_R	Abbreviated $N_{R,b}$ when there is no ambiguity	
$N_{B,b}$	Number (Density) of blue agents in block b	
N_B	Abbreviated $N_{B,b}$ when there is no ambiguity	
z_b	Normalization constant for P_b	DFFT
$v_{R,b}$	Vexation for red agents in block b	
$v_{B,b}$	Vexation for blue agents in block b	
f	Frustration	
H_b	Headache function for block b	
P	Probability distribution of states of the entire city	
$P_{b \rightarrow b'}$	Probability of transition of an agent from block b to b'	
$v_{R,b \rightarrow b'}$	Number (Density) flow rate for red agent from block b to b'	
$v_{B,b \rightarrow b'}$	Number (Density) flow rate for blue agent from block b to b'	
μ_R	Red agent potential in a city	
μ_B	Blue agent potential in a city	

$$P_{\text{Schelling}} = \begin{cases} \dfrac{1}{1+e^{-\Delta(U_R^{\text{so}}+U_R^{\text{sp}})}} & \text{if agent is red} \\[2ex] \dfrac{1}{1+e^{-\Delta(U_B^{\text{so}}+U_B^{\text{sp}})}} & \text{if agent is blue} \end{cases}, \qquad (1)$$

where Δ denotes the change in utilities due to the proposed move [28] so that agents are more likely to move if the total utility increases. The social and spatial utility functions are defined in Figs. 2b,c and 2d,e respectively. In this particular case, we

define the social utility to be proportional to the number of 8-connected neighbors (dashed box in Fig. 2a) that are of the same type (more complex social utility functions are considered in SI Section S8). This dependence is illustrated by the monotonicity of the utility functions in the rows and columns in Fig. 2b and c respectively. We set the spatial utilities as shown in Fig. 2d and e, where red agents prefer the West side of the city and blue agents prefer the South side of the city. We use these simulations to generate the data throughout this paper.

From these simulations we obtain coarse-grained data (i.e., data at some small-area organizational unit) of agent densities and their steady-state joint probability distributions. In particular, we run an ensemble of Schelling simulations until they reach a steady state, where the probability of a block having a particular density no longer changes over time (See SI Section S1 for a discussion of how an ensemble may be obtained for real-world data). A sample steady-state configuration for the ensemble is shown in Fig. 2f. We coarse-grain the Schelling lattice grid into 25 blocks (outlined in the figure by thick lines), such that there are 144 sites in each block, and record the total number of red agents $N_{R,b}$ and blue agents $N_{B,b}$ in each block b. Since all the blocks have the same area, $N_{R,b}$ and $N_{B,b}$ indicate the local densities. By sampling the different steady-state configurations (denoted in the figure by the stack and ellipses) we measure the joint local probability distribution of agent densities for each block $P_b(N_{R,b}, N_{B,b})$. For simplicity, we will abbreviate $N_{R,b}$ and $N_{B,b}$ as N_R and N_B when there is no ambiguity. We show the steady-state joint probability distributions for the North East (NE), South West (SW) and South East (SE) blocks in fig. 2g. As a result of spatial utility, the SE block is likely to be occupied by a high density of blue agents, while the NE block is likely to be occupied by a low density of red and blue agents. The SW block is occupied by high densities of agents with a wide distribution of red and blue agent compositions. This wide distribution reflects the tendency of the red and blue agents to segregate due to social utility.

Density-functional fluctuation theory

Single-group Density-Functional Fluctuation Theory posits that, by observing the steady-state probability distribution in a block (a small-area organizational unit), one can extract information about the location-dependent preferences and social interactions for a single type of agent [30]. In particular, by observing the means of the distributions, it is possible to determine the agent preference for each block, while the shapes of the distributions provide information about social preference. For example, a Poisson-like distribution for the density in a block indicates no social interactions, a narrowly peaked distribution indicates strong repulsion, and a bimodal distribution indicates strong attractive interactions. [30] developed a theory that captures these behaviors through a block-dependent steady-state probability distribution of the form

$$P_b(N) = z_b^{-1} \frac{1}{N!} \exp[-v_b N - f(N)], \tag{2}$$

Fig. 2 Schelling-type simulation and steady-state data. a Top-left corner of the Schelling lattice grid ▶ with 1000 red and 1000 blue agents. At each step in time, an agent and an empty cell are randomly chosen, and the agent will make probabilistic move to the empty cell. In this example, a red agent is chosen to move to a randomly chosen empty cell. The 8-connected neighborhood of the red agent and empty cell are shown as dashed boxes. **b** Social Utility for red agents is defined by $U_R^{\mathrm{so}}(N_R^{\mathrm{ne}}, N_B^{\mathrm{ne}}) = 0.4 \cdot N_R^{\mathrm{ne}}$, where N_R^{ne} and N_B^{ne} are the number of 8-connected red and blue neighbors respectively. For the red agent in **a**, the change in social utility due to the proposed move is given by $U_R^{\mathrm{so}}(5, 1) - U_R^{\mathrm{so}}(1, 6) = +1.6$, making this move favorable by social preferences. **c** Social Utility for blue agents is defined by $U_B^{\mathrm{so}}(N_R^{\mathrm{ne}}, N_B^{\mathrm{ne}}) = 0.4 \cdot N_B^{\mathrm{ne}}$. **d** Spatial Utility for red agents $U_R^{\mathrm{sp}}(x)$ is a function of location x, that decreases linearly in the horizontal direction. The change in spatial utility for the red agent in **a** for the proposed move is $\Delta U_R^{\mathrm{sp}} \approx +0.17$, making this move favorable by spatial preferences. So, according to Eq. (1), the red agent has a 85% chance of moving. **e** Spatial Utility for blue agents $U_B^{\mathrm{sp}}(x)$ is a function of location x, that decreases linearly in the vertical direction. **f** A sample steady-state configuration of our simulation after reaching steady state (> 10000 steps). We divide the Schelling lattice grid into 25 blocks with 144 sites each, three of which are shaded and labeled as 'NE', 'SW', and 'SE' for reference. To obtain a collection of steady-state configurations (red stack and red ellipses), we run an ensemble of Schelling simulations. **g** From the collection of steady-state configurations, one can observe the steady-state joint probability distribution of observing a given agent densities for each block. Distributions for blocks 'NE', 'SW', 'SE' are shown

where z_b^{-1} is a normalization constant; v_b is defined as the "vexation" and is block b specific; and $f(N)$ is defined as the "frustration" and is a block-independent function of local densities. Since we presently consider blocks of the same size, the number of agents N is commensurate with the density. When $f(N)$ is zero, the distribution is Poisson and the mean is proportional to $\exp[-v_b]$, indicating agents avoid blocks with high v_b. The deviation of the distribution from Poisson is captured by the function $f(N)$ that depends only on the density of agents. When $f(N)$ is concave up (e.g., P_b is narrowly peaked), agents in the city disperse. When $f(N)$ is concave down (e.g., P_b is bimodal) agents in the city aggregate. Importantly, $f(N)$ is non-parametric and thus it captures the effective social interactions in an entirely data-driven fashion. Thus, frustration and vexations respectively capture effective social and spatial interactions directly from the data and do not require *a priori* knowledge of agent preferences. In previous work, it was shown that this functional form for the probability distribution arises mathematically from the detailed balance condition for the equilibrium in a broad class of agent-based models. In fact, DFFT has proven remarkably accurate for predicting crowd distributions in not only modelled but also living systems [30].

Furthermore, it has been demonstrated that such a model arises for real-world cities by applying the principle of maximum entropy (MaxEnt) to U.S. Census data [2]. Specifically, by employing the size-invariance and organizational equivalence properties, which all common segregation indices of evenness and exposure satisfy, [2] are able to transform Census data into a probability density function with an analogous 2-group DFFT form,

$$P_b(N_R, N_B) = z_b^{-1} \frac{1}{N_R! N_B! (s - N_R - N_B)!} \exp[-v_{R,b} N_R - v_{B,b} N_B - f(N_R, N_B)],$$

$$(3)$$
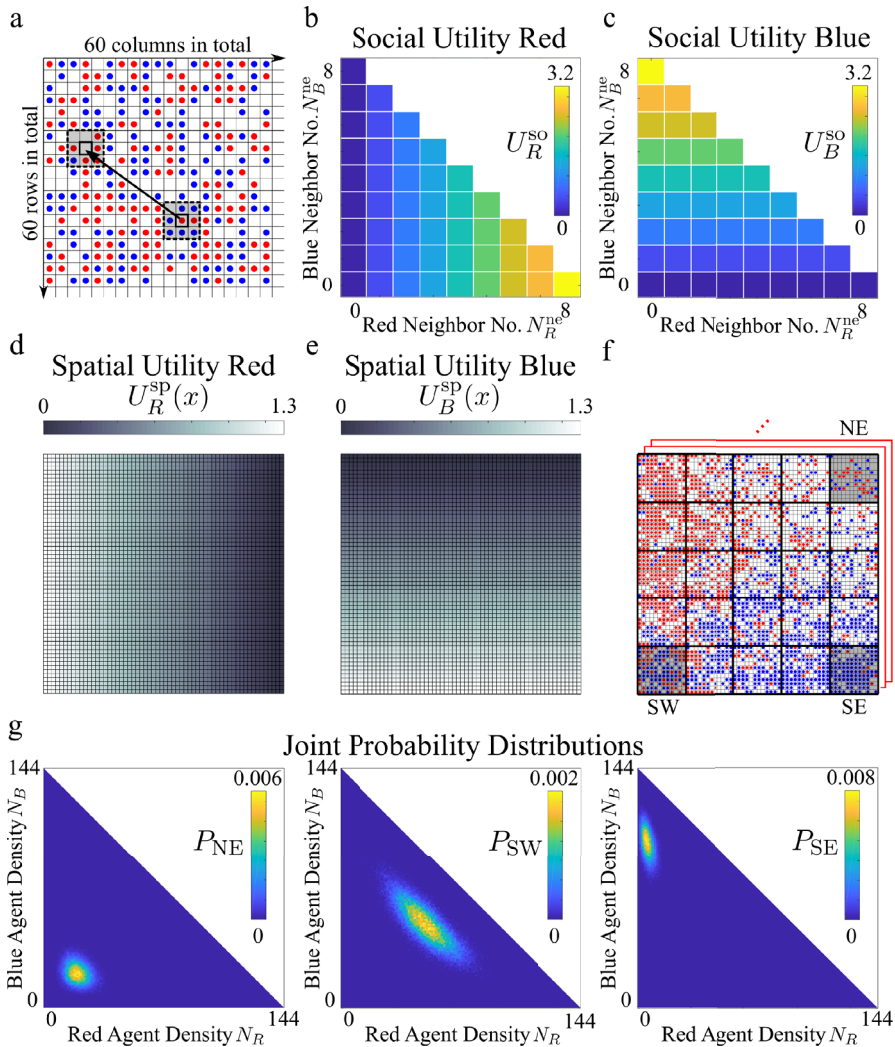
where z_b^{-1} is, again, a normalization constant; $v_{R,b}$ and $v_{B,b}$ are the block-dependent vexations for the Red and Blue agents; s is the total number of sites within a block; and $f(N_R, N_B)$ is the block-independent frustration that is a function of local densities of Red and Blue agents. The $1/(s - N_R - N_B)!$ term arises to account for blocks being of finite size in the work of [2], where neighborhoods have a known total population, but not [30] where there is no set limit to the number of agents in a block. To be precise, Eq. (3) corresponds to the 'three-group case' in the analysis performed by [2] where an empty site is effectively considered as a third type of agent. Despite this, we refer to this as 2-group DFFT due to the presence of only two 'real' agent types.

As in the single-group case, agents avoid blocks with a high vexation. However, the frustration is now interpreted as a surface that captures the social interactions for two types of agents. The concavities of curves on this surface reflect the social preferences for having greater or fewer agents of a particular type, and the frustration itself can be considered a multigroup, non-parametric, functional measure of segregation [2].

To verify that Eq. (3) can indeed be used to fit the Schelling model steady-state data (Fig. 2g), we first rearrange Eq. (3) to obtain

$$-\ln[N_R!N_B!(s - N_R - N_B)!P_b(N_R, N_B)] = f(N_R, N_B) + v_{R,b}N_R + v_{B,b}N_B + c_b,$$
(4)

where $c_b = \ln(z_b)$ is a normalization constant. The left-hand side (abbr. LHS) of Eq. (4) is determined by our observed probability P_b (Fig. 2g), and is plotted for three sample blocks in Fig. 3a. Next, we use a Maximum Likelihood Estimation algorithm to infer the frustration and vexations that best fit the data and plot these in Fig. 3b. We find that the fits are remarkably accurate as illustrated by the small errors, differences between left-hand side and right-hand side (abbr. RHS) of Eq. (4), shown for three sample blocks in Fig. 3c. When comparing the observed joint probability distributions (Fig. 2g) with the distributions modeled by Eq. (3) using the extracted frustration and vexations (Fig. 3b), we observe a mean absolute percentage error of 14% for joint densities with at least 10 observations. As we now show, the extracted frustration and vexations can next be used to predict how small-area populations will redistribute in response to regional population changes.

Predicting time evolution

To generate a regional population change in the simulation data, we abruptly switch 350 randomly chosen red agents in the north half of the Schelling lattice into blue agents. Although such a change is unrealistic, it results in an extreme small-area time evolution and therefore is harder to forecast accurately than neighborhood changes arising from more modest regional changes. We then record the evolution of an ensemble of cities undergoing this regional change (illustrated by the stack and ellipses), going from the initial altered state at $t = 0$ to the new steady state at $t \rightarrow \infty$ (Fig. 4a). As before, we coarse grain these data at the block level to extract the density of red and blue agents at each point in time. The challenge is then to predict the evolution of the densities at the block scale using DFFT parameters extracted from the initial steady state data and knowledge of the regional-scale demographic change.

Time-dependent DFFT model (Kohn-Sham TD-DFFT)

To accurately predict the time evolution of small-area populations, we introduce a Time-Dependent DFFT model. This method parallels the adiabatic approximation
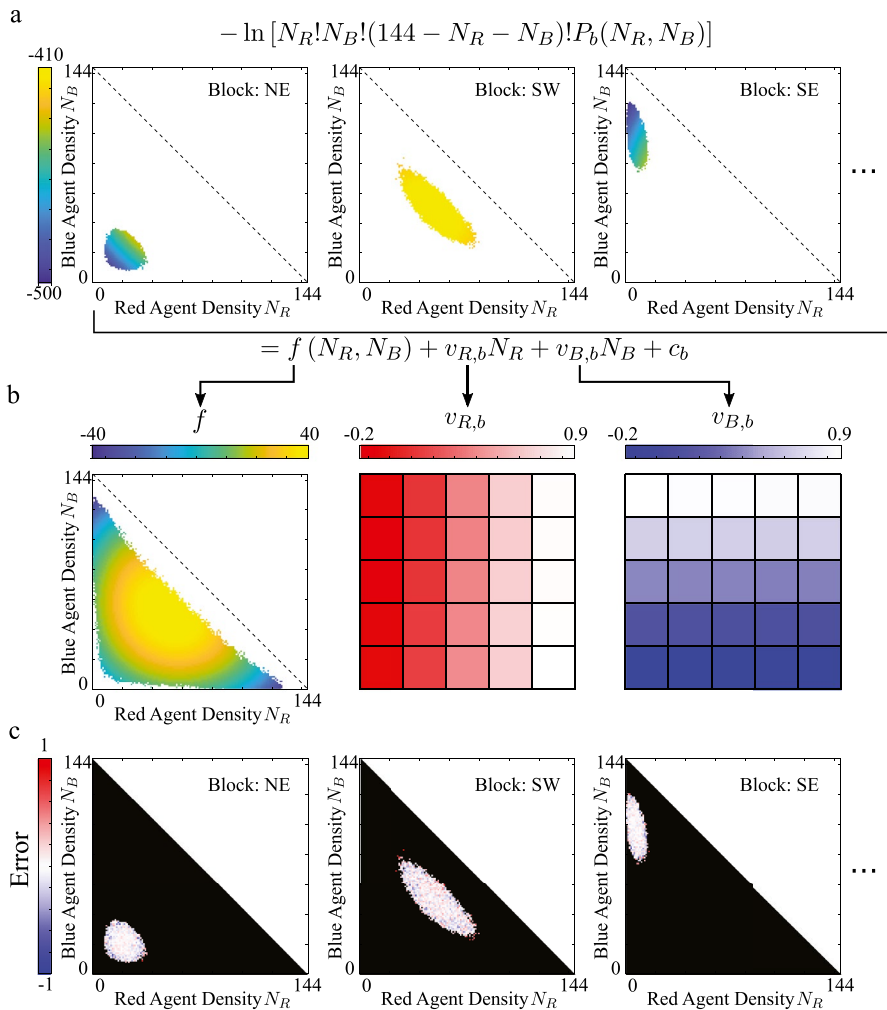
Fig. 3 **Extracting effective social and spatial preferences a** The LHS of Eq. (4) is determined by our observed probability P_b (Fig. 2g), and is plotted for blocks 'NE', 'SW', 'SE'. We only keep data for cases where more than 10 observations are recorded for a particular agent combination. **b** Using Maximum Likelihood Estimation, we fit each of the 25 LHS surfaces by a block-independent surface called "frustration" together with a block-dependent planar shift $v_{R,b}N_R + v_{B,b}N_B + c_b$, where $v_{R,b}$ and $v_{B,b}$ are block-dependent constants called "vexations", and c_b is a block-dependent normalization constant. Frustration describes social preference, while vexations describe spatial preference. **c** The errors of the fit in **b** are determined by the difference between the right hand and left hand sides of Eq. (4), for the NE, SW, and SE blocks and demonstrate very good agreement (generally in the range of only ±1, which is two orders of magnitude smaller than the variation in the RHS of Eq. (4)). We measure mean absolute error (MAE) to be 0.12, 0.15, and 0.13 (out of a dynamic range of ~100) for blocks 'NE', 'SW', and 'SE', respectively. We measure Pearson correlations to be 0.9999, 0.9975, and 0.9999 for blocks 'NE', 'SW', and 'SE', respectively

found in Kohn-Sham Time-Dependent Density-Functional Theory (TD-DFT). The strength of the TD-DFFT model is that it relies solely on the block-scale (small-area)
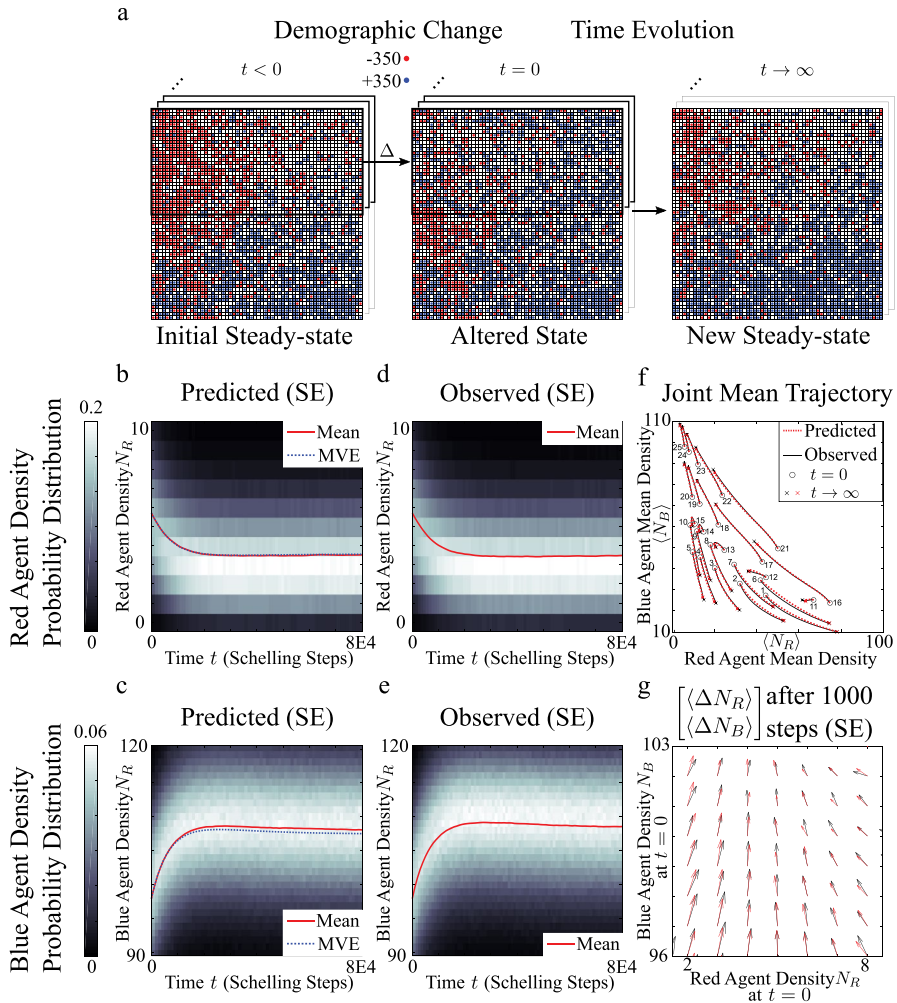
Fig. 4 Predicting time evolution a Starting from a steady-state configuration, at $t = 0$, we introduce our demographic change by abruptly switching 350 randomly chosen red agents on the north side of the Schelling lattice into blue agents to obtain an altered state. The system is then evolved according to the Schelling model. The above procedure is repeated over an ensemble of Schelling simulations (shown as stacks and ellipses). **b-c** Predicted time evolution of the probability distribution for observing red or blue agents for the 'SE' block using the TD-DFFT model. Note that the distribution for red agents is skewed away from the mean towards more segregated values. The MVE predictions agree well with the Mean of the TD-DFFT Model for both types of agent. **d-e** Observed time evolution of the probability distribution for observing red or blue agents for the 'SE' block from the Schelling simulation. For $0 < t < 20000$, we measure mean absolute error (MAE) to be 0.10 (out of a dynamic range of ~3) and 0.19 (out of a dynamic range of ~10) for mean values of red and blue agents, respectively; we measure Pearson correlations to be 0.9981 and 0.9983 for mean values of red and blue agents, respectively. **f** Observed versus predicted Joint-mean density trajectories for all blocks (counted left-to-right then top-to-bottom, in normal English reading order). Blocks 13-15 show interesting trajectories, which the TD-DFFT model predicts well. **g** Observed versus predicted average changes in number of agents in block 'SE' after 1000 Schelling steps for various initial numbers of agents. We note that a calibration factor is necessary to match the time scales between the density based model predictions and the Schelling simulations

densities and the inferred density-dependent spatial and social preferences to construct an *effective* agent-based model. These coarse-grained preferences can be combined to form block-level "Headache" functions,

$$H_b(N_R, N_B) \equiv v_{R,b} N_R + v_{B,b} N_B + f(N_R, N_B). \tag{5}$$

The operational steps of the TD-DFFT model are then as follows. First, at every time step, an agent is randomly selected and a target block is chosen with a weight proportional to the number of empty spaces it has. The agent then moves from its current block b to the chosen block b' with the probability

$$P_{b \to b'} = \frac{1}{1 + e^{\Delta H_b + \Delta H_{b'}}}, \tag{6}$$

where Δ represents the change due to the proposed move.

This Kohn-Sham version of TD-DFFT has two advantages over traditional (*e.g.,* Schelling) agent-based models that presume agent interactions. First, the effective agent utilities are *inferred* from observational data of small-area populations. Second, TD-DFFT can be implemented more efficiently than a Schelling-type agent-based model because, in TD-DFFT, the city is represented entirely by small-area population counts, making it unnecessary to represent the city at the individual-level. For instance, it is possible to start with a particular city-scale state and numerically simulate agent migrations using Eq. (6) to generate a realization for the time-evolution of block-scale densities. Additionally, it is also possible to predict the joint probability distribution of the ensemble directly, without an agent-based simulation, through a master equation approach (See SI Section S2). Crucially, as we will demonstrate in the following sections, these TD-DFFT approaches align well with the time-evolution of density changes found directly from the underlying agent-based model once a constant time-scale is introduced to align the average rate of TD-DFFT migrations with the observed average migration rate (See SI Section S2.3 for a more detailed discussion of this time-scale).

Mean Value Equation (Hohenberg-Kohn TD-DFFT)

For cases where additional computational simplicity is necessary—especially when the number of city-scale states becomes intractable—we develop a simplified mean-value approach of the TD-DFFT model described above. In particular, we note that real-world cities can have thousands of neighborhoods and millions of people, and that the number of feasible city-scale states scales steeply with the number of blocks and the number of individuals within blocks. The salient assumption of the mean-value approach is then to reduce the distribution of possible densities for each block to its mean density, allowing for the average city behavior to be represented by a number of variables equivalent to the number of blocks multiplied by the number of agent types.

To apply the Mean Value Equation (MVE) (See derivation in SI Section S2.2.), it is necessary that the probability distributions of small-area densities are single-peaked, which is indeed the case here, but may not be the case in cases of very

strong segregation. In particular, MVE assumes that the behavior of a small-area unit is well represented by the behavior of its expected (average) composition. If a neighborhood's population could be either dominated by one group or the other with some probability, then treating the neighborhood as a uniform mixture of the two groups, as MVE would do, will be a poor approximation. Under conditions where this is not a concern, then the MVE can be applied as a very efficient computational approach, and the time-evolution of the TD-DFFT model can be approximated using

$$
\begin{cases}
\dfrac{\mathrm{d}}{\mathrm{d}t}\overline{N_{R,b}} = \displaystyle\sum_{b'\neq b}\left(v_{R,b'\to b} - v_{R,b\to b'}\right) \\[2ex]
\dfrac{\mathrm{d}}{\mathrm{d}t}\overline{N_{B,b}} = \displaystyle\sum_{b'\neq b}\left(v_{B,b'\to b} - v_{B,b\to b'}\right)
\end{cases},
\tag{7}
$$

where $\overline{N_{R,b}}$ and $\overline{N_{R,b}}$ denote the mean red and blue agent density for block b, respectively. $v_{R,b\to b'}$ and $v_{B,b\to b'}$ are the density flow rates for red and blue agents from block b to block b', calculated from

$$
\begin{cases}
v_{R,b\to b'} \approx \dfrac{\overline{N_{R,b}}}{N_{\mathrm{tot}}} \cdot \dfrac{s - \overline{N_{R,b'}} - \overline{N_{B,b'}}}{s_{\mathrm{tot}} - N_{\mathrm{tot}}} \cdot P_{b\to b'} \\[3ex]
v_{B,b\to b'} \approx \dfrac{\overline{N_{B,b}}}{N_{\mathrm{tot}}} \cdot \dfrac{s - \overline{N_{R,b'}} - \overline{N_{B,b'}}}{s_{\mathrm{tot}} - N_{\mathrm{tot}}} \cdot P_{b\to b'}
\end{cases}.
\tag{8}
$$

Eq. (7) says that the rate of change in the mean density of agents in block b is given by the sum of inflow rates from all other blocks $b' \neq b$ into block b, minus the sum of outflow rates from block b into all other blocks $b' \neq b$. We can also easily interpret the flow rate approximations in Eq. (8) following the rules of the TD-DFFT Model (Sect. 4.1): The first term in the product represents the probability of choosing the corresponding type of agent in block b, where the denominator N_{tot} is the total number of agents in the city; the second term in the product represents the probability of choosing an empty cell in block b', where the denominator $s_{\mathrm{tot}} - N_{\mathrm{tot}}$ is the total number of empty cells in the city; the third term corresponds to the probability of transition defined in Eq. (6). Note that this MVE might fail to capture the behavior of the TD-DFFT Model with extreme segregation, when Eq. (7) exhibits 'bifurcation behavior' [19, 55] (See SI Section S3). Since the MVE deals with average numbers in each block, it corresponds more closely to a Hohenberg-Kohn TD-DFT [20].

Results

Overall, we obtain excellent agreement between the TD-DFFT model/MVE predictions and the simulated time-dependent Schelling data. As an example, we compare the predicted (Figs. 4b,c) and observed (Figs. 4d,e) time evolution of the probability distribution of red (N_R) and blue (N_B) agents for the South East (SE) block (and all

other blocks in SI Section S4). We find that the model accurately captures the trends in the means, Pearson correlation with observations >0.99. as well as the skews in the distributions about the means. In particular, the MVE (blue dotted lines in Figs. 4b,c) accurately tracks the mean values of the TD-DFFT model (red lines in Figs. 4b,c). Additionally, we compare the trajectory of the joint means for all the blocks in Fig. 4f. Once again, we find excellent agreement throughout the entire trajectory—even when the evolution is non-monotonic as is the case for blocks 12 through 15. Finally, given an initial joint density (N_R, N_B) for a particular block, we are able to predict the joint probability distribution for the block after 1000 Schelling steps. We plot the observed average change in the joint density for each initial condition for the SE block in Fig. 4g. These predictions, indicated by the red arrows, are compared with the observed data, indicated by the black arrows. Once again, we observe excellent agreement between the predictions of the TD-DFFT model and the small-area population data. It is this capacity to model the step-wise evolution that allows the TD-DFFT model to accurately track the time dependent trajectories of the Joint Means as shown in Fig. 4f. Finally, we have conducted similar studies involving the same demographic change for more complicated Social and Spatial utility functions and have obtained TD-DFFT predictions of similar fidelity (SI Section S8).

In part, the reason the Time-dependent DFFT model is able to accurately describe the evolution of the Schelling data after a demographic change is that it is also a type of agent-based model. There are, however, a number of important distinctions. First, the TD-DFFT model requires only coarse-grained data, which means it can work with the substantially more accessible small-area data as opposed to requiring data at the individual scale. Details of the Schelling simulation such as the lattice grid structure, the 8-connected neighbors, and empty spots are all averaged over to obtain the density in each block. As such, the TD-DFFT model keeps only the essential information necessary to make predictions about the density. Second, the TD-DFFT model relies on empirically extracted parameters and so does not require that we impose specific agent preferences governing the system evolution. Instead, it determines the necessary preferences from the original steady-state data in order to make predictions. These distinctions can lead to discrepancies in certain extreme conditions. For instance, when the block-size is very small (e.g., 4 by 4 cells) the interaction of agents between neighboring blocks will greatly affect the dynamics (See SI Section S5 for discussion of this discrepancy).

Predicting the new steady-state

To predict the new steady-state joint probability distributions, $P_b(N_R, N_B, t \rightarrow \infty)$, we can either run the TD-DFFT model until it reaches a steady state or calculate the new distribution analytically using DFFT. To analytically predict the new distribution, we take advantage of the fact that the social and spatial preferences of the individuals remain the same following the regional demographic change in the total number of each agent type. Since these social and spatial preferences manifest

themselves at the coarse-grained level as differences in the headache functions of two blocks, $\Delta H_b + \Delta H'_b$, we can rewrite the headache function as

$$H_b \longrightarrow H_b - \mu_R N_R - \mu_B N_B \tag{9}$$

without affecting these preferences. In other words, $\Delta H_b + \Delta H'_b$ remains unaffected following this transformation. Here, μ_R and μ_B are block-independent constants called 'agent potentials' (directly analogous to chemical potentials in statistical physics and density-functional theory) that tune the expected total number of each type of agent over all the blocks. Such a change was also shown by [2] to result from minimizing the cross-entropy under the constraint of a known city-wide change in the number of agents. To determine the agent potential constants, we modify the exponent in Eq. 3 and use Newton's method to converge on values for μ_R and μ_B so
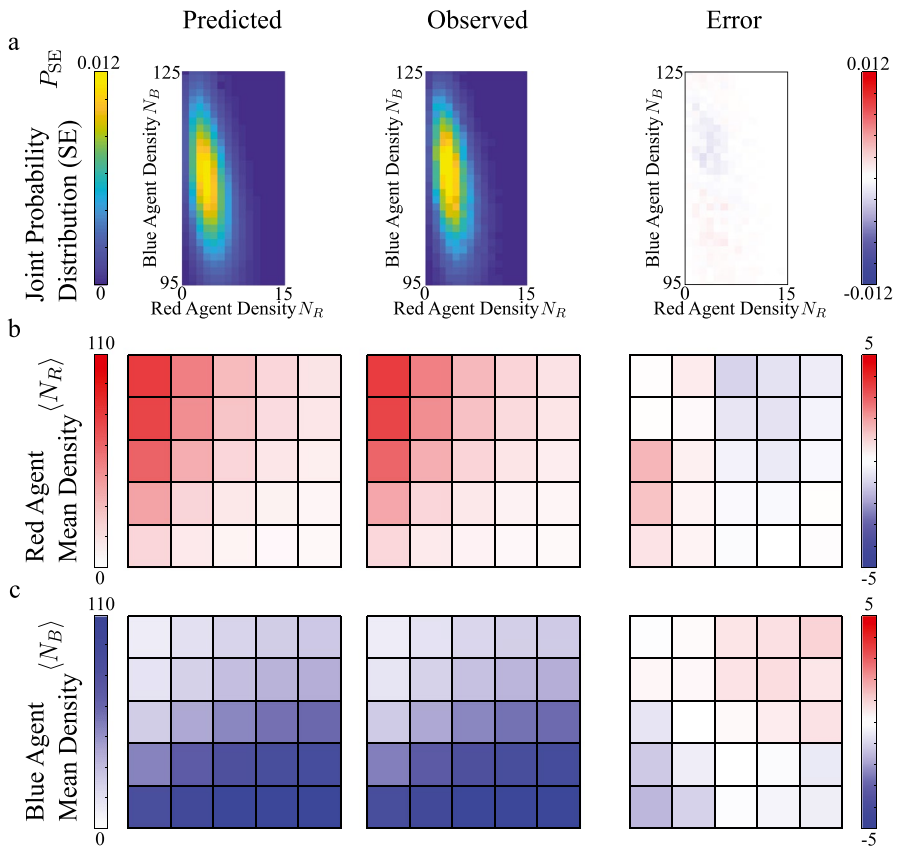
Fig. 5 Analytic prediction of new steady state. a Predicted versus observed new steady-state joint probability distribution for block 'SE'. **b** Predicted versus observed mean densities of red agents for all blocks, each with total capacity 144. We measure a Pearson correlations of 0.9997 and mean absolute error (MAE) of 0.42 (out of a dynamic range of ~100). **c** Predicted versus observed mean densities of blue agents for all blocks, each with total capacity 144. We measure a Pearson correlation of 0.9999 and MAE value of 0.50 (out of a dynamic range of ~100)

that the means of the probability distribution, when averaged over all the blocks, equals the new mean for the entire city following the demographic change. We compare the predictions for the joint probability distributions, as well as the mean of the agent densities for each block with the observed data from the Schelling model in Fig. 5 (See SI Section S4 for results for all blocks). We find excellent agreement between the predictions and the simulated small-are population data. Importantly, this analytic approach indeed arrives at the same new steady state distribution that our TD-DFFT model predicts (SI Section 2.1).

Implications and future directions

The ability of DFFT to accurately predict demographic changes in the Schelling model suggests that it may improve upon current methods for small-area population forecasting. Unlike bottom-up agent-based approaches that postulate specific agent preferences, DFFT empirically extracts these preferences from observations. Additionally, in contrast to top-down data-driven approaches that only extract descriptive measures of segregation, DFFT uses more detailed measures that allow for the forecasting of population dynamics. Importantly, this framework could easily be extended to include an even greater number of agent types (potentially due to the presence of multiple racial groups, or parsing racial groups by other demographic characteristics into further subgroups) as the process of including agent-type-specific vexations and a multidimensional frustration is straightforward.

We also anticipate a pathway towards modeling other demographic changes, such as changes in social or spatial preferences. In these particular cases, it is necessary to know how to map such changes to the DFFT functions, which can be achieved in a number of ways. First, it may be possible to estimate the change in the DFFT functions based on certain sociological or demographic factors. Second, even if we had no *a priori* knowledge of this change, we could still use intermediate time points in the evolution of the demographic data to adjust the vexations and frustration. For example, after a certain number of Schelling time steps we would modify our DFFT functions so that we get the best agreement between the dynamics of the predicted and observed density changes for the agents (Fig. 4g). This suggests that DFFT can be a rather flexible approach for making predictions in a broad range of demographic conditions. Finally, to the extent that agent-based approaches like the Schelling model inform trends in demographic patterns, it may be possible to apply TD-DFFT to real data. For example, decennial U.S. Census data provides block-level counts of the number of people by race and ethnicity. Given such data, DFFT could now be used not only to quantify segregation throughout the country [2] but also to forecast neighborhood-level changes in a manner which directly incorporates the observable effects of segregation.

Author Contributions T.A.A. developed initial theoretical extensions to DFFT work including multi-group and time-dependent systems. Y.A.K. and I.C. proposed initial application of DFFT onto demographic systems. M.H. provided context of proposed methods within the broader field of demography. Y.C., Y.A.K., B.B., T.A.A. and I.C. developed and refined multi-group and time-dependent applications of DFFT onto simulated data. Y.C. implemented all simulations, performed all analyses of data, and created all figures. Y.A.K. wrote code for statistical extraction of parameters from data. Y.C., Y.A.K., B.B., T.A.A. and I.C. wrote the manuscript with all authors contributing. Y.C., T.A.A and B.B. wrote the SI with all authors contributing. I.C. mentored Y.C. and Y.A.K.. T.A.A. mentored Y.C. and B.B..

Data Availability Statement The code used to to run the simulations and perform statistical analyses, as well as datasets generated by the simulation, are available at https://github.com/yunuskink/DFFT-Schelling-model.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Anderson, L. M., Charles, J. S., Fullilove, M. T., et al. (2003). Providing affordable family housing and reducing residential segregation by income: A systematic review. *American Journal of Preventive Medicine, 24*(3), 47–67.
2. Barron, B., Kinkhabwala, Y.A., Hess, C., et al (2022) Extending the use of information theory in segregation analyses to construct comprehensive models of segregation. arXiv preprint arXiv:2212.06980
3. Benenson, I., Hatna, E., & Or, E. (2009). From schelling to spatially explicit modeling of urban ethnic and economic residential dynamics. *Sociological Methods & Research, 37*(4), 463–497.
4. Boke-Olén, N., Abdi, A. M., Hall, O., et al. (2017). High-resolution African population projections from radiative forcing and socio-economic models, 2000 to 2100. *Scientific Data, 4*(1), 1–9.
5. Bruch, E. E., & Mare, R. D. (2006). Neighborhood choice and neighborhood change. *American Journal of Sociology, 112*(3), 667–709.
6. Bruch, E. E., & Mare, R. D. (2009). Preferences and pathways to segregation: Reply to van de rijt, siegel, and macy. *American Journal of Sociology, 114*(4), 1181–1198.
7. Card, D., Mas, A., & Rothstein, J. (2006). Tipping and the dynamics of segregation in neighborhoods and schools
8. Chi, G. (2009). Can knowledge improve population forecasts at subcounty levels? *Demography, 46*(2), 405–427.
9. Chyn, E., & Katz, L. F. (2021). Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives, 35*(4), 197–222.

10. Clark, W. A. (1986). Residential segregation in American cities: A review and interpretation. *Population Research and Policy Review, 5*(2), 95–127.

11. Clark, W. A., & Fossett, M. (2008). Understanding the social context of the schelling segregation model. *Proceedings of the National Academy of Sciences, 105*(11), 4109–4114.

12. Echenique, F., & Fryer, R. G., Jr. (2007). A measure of segregation based on social interactions. *The Quarterly Journal of Economics, 122*(2), 441–485.

13. Ellis, M., Wright, R., Fiorio, L., et al. (2018). Predicting neighborhood racial change in large us metropolitan areas, 1990–2010. *Environment and Planning B: Urban Analytics and City Science, 45*(6), 1022–1037.

14. Foot, D. K., & Milne, W. J. (1984). Net migration estimation in an extended, multiregional gravity model. *Journal of Regional Science, 24*(1), 119–133.

15. Fowler, C. S. (2016). Segregation as a multiscalar phenomenon and its implications for neighborhood-scale research: The case of south seattle 1990–2010. *Urban Geography, 37*(1), 1–25.

16. Freeman, L. (2009). Neighbourhood diversity, metropolitan segregation and gentrification: What are the links in the us? *Urban Studies, 46*(10), 2079–2101.

17. Grauwin, S., Goffette-Nagot, F., & Jensen, P. (2012). Dynamic models of residential segregation: An analytical solution. *Journal of Public Economics, 96*(1–2), 124–141.

18. Grogger, J., & Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics, 95*(1), 42–57.

19. Haag, G. (2017). *Modelling with the Master Equation: Solution Methods and Applications in Social and Natural Sciences*. Cham, Switzerland: Springer International Publishing.

20. Hohenberg, P., & Kohn, W. (1964). *Inhomogeneous electron gas. Physical review, 136*(3B), B864.

21. Homandberg, L., Hall, M. (2024). Spatial dynamics of white flight: A novel approach. PAA annual meeting 2024

22. Humphreys, J. S. (1998). Delimiting 'rural': implications of an agreed 'rurality'index for healthcare planning and resource allocation. *Australian Journal of Rural Health, 6*(4), 212–216.

23. Karemera, D., Oguledo, V. I., & Davis, B. (2000). A gravity model analysis of international migration to north america. *Applied Economics, 32*(13), 1745–1755.

24. Keyfitz, N., & Caswell, H. (2005). *Applied mathematical demography* (Vol. 47). New York, United States: Springer.

25. Kim, K., & Cohen, J. E. (2010). Determinants of international migration flows to and from industrialized countries: A panel data approach beyond gravity. *International Migration Review, 44*(4), 899–932.

26. Land, K. C., Yang, Y., & Zeng, Y. (2005). *Mathematical demography*. New York, United States: Springer.

27. Mason, A. (1996). Population and housing. *Population Research and Policy Review, 15*(5–6), 419–435.

28. McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. Institute of Urban and Regional Development

29. McKee, J. J., Rose, A. N., Bright, E. A., et al. (2015). Locally adaptive, spatially explicit projection of us population for 2030 and 2050. *Proceedings of the National Academy of Sciences, 112*(5), 1344–1349.

30. Méndez-Valderrama, J. F., Kinkhabwala, Y. A., Silver, J., et al. (2018). Density-functional fluctuation theory of crowds. *Nature Communications, 9*(1), 3538.

31. Mora, R., & Ruiz-Castillo, J. (2011). Entropy-based segregation indices. *Sociological Methodology, 41*(1), 159–194.

32. Oka, M., & Wong, D. W. (2014). Capturing the two dimensions of residential segregation at the neighborhood level for health research. *Frontiers in Public Health, 2*, 118.

33. Oka, M., & Wong, D. W. (2016). Spatializing area-based measures of neighborhood characteristics for multilevel regression analyses: An areal median filtering approach. *Journal of Urban Health, 93*(3), 551–571.

34. Park, Y. M., & Kwan, M. P. (2018). Beyond residential segregation: A spatiotemporal approach to examining multi-contextual segregation. *Computers, Environment and Urban Systems, 71*, 98–108.

35. Parrott, J.T.B., Carnevale, S. (1997) Locating fire station sites: The response lime component. Demographics: A Casebook for Business and Government 904:203

36. Poot, J., Alimi, O., Cameron, M.P., et al (2016). The gravity model of migration: the successful comeback of an ageing superstar in regional science. IZA discussion paper

37. Preston, S., Heuveline, P., & Guillot, M. (2000). *Demography: Measuring and Modeling Population Processes*. New Jersey, United States: Wiley-Blackwell.

38. Ramos, R., & Suriñach, J. (2017). A gravity model of migration between the enc and the eu. *Tijdschrift voor Economische en Sociale Geografie, 108*(1), 21–35.

39. Reardon, S. F. (2006). A conceptual framework for measuring segregation and its association with population outcomes. *Methods in social epidemiology, 1*(169), 169–192.

40. Reardon, S. F., & Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological Methodology, 32*(1), 33–67.

41. Reardon, S. F., & O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology, 34*(1), 121–162.

42. Van de Rijt, A., Siegel, D., & Macy, M. (2009). Neighborhood chance and neighborhood change: A comment on bruch and mare. *American Journal of Sociology, 114*(4), 1166–1180.

43. Rogers, A. (2008). Demographic modeling of the geography of migration and population: A multi-regional perspective. *Geographical Analysis, 40*(3), 276–296.

44. Rowland, D.T. (2003). Demographic methods and concepts. OUP Catalogue

45. Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1*(2), 143–186.

46. Siegel, J. S., & Swanson, D. A. (2004). *The Methods and Materials of Demography*. Bingley, United Kingdom: Emerald Group Publishing Limited.

47. Smith, S. K., Tayman, J., & Swanson, D. A. (2013). *A practitioner's guide to state and local population projections*. New York, United States: Springer.

48. Spaiser, V., Hedström, P., Ranganathan, S., et al. (2018). Identifying complex dynamics in social systems: A new methodological approach applied to study school segregation. *Sociological Methods & Research, 47*(2), 103–135.

49. Swanson, D. A., Hough, G., Rodriguez, J. A., et al. (1998). K-12 enrollment forecasting: merging methods and judgment. *ERS spectrum, 16*(4), 24–31.

50. Tita, G. E., Petras, T. L., & Greenbaum, R. T. (2006). Crime and residential choice: A neighborhood level analysis of the impact of crime on housing prices. *Journal of Quantitative Criminology, 22*, 299–317.

51. Vinković, D., & Kirman, A. (2006). A physical analogue of the schelling model. *Proceedings of the National Academy of Sciences, 103*(51), 19261–19265.

52. Wachter, K. W. (2014). *Essential demographic methods*. Massachusetts, United States: Harvard University Press.

53. Weidlich, W. (2006). *Sociodynamics: A systematic approach to mathematical modelling in the social sciences*. Massachusetts, United States: Courier Corporation.

54. Weidlich, W., & Haag, G. (1988). *Interregional migration: dynamic theory and comparative analysis* (Vol. 4). New York, United States: Springer.

55. Weidlich, W., & Haag, G. (2012). *Concepts and models of a quantitative sociology: the dynamics of interacting populations* (Vol. 14). New York, United States: Springer Science & Business Media.

56. Weiner, E. (2016). *Urban transportation planning in the United States: history, policy, and practice*. New York, United States: Springer.

57. Wellman, G. C. (2014). Transportation apartheid: the role of transportation policy in societal inequality. *Public Works Management & Policy, 19*(4), 334–339.

58. White, M. J., Kim, A. H., & Glick, J. E. (2005). Mapping social distance: Ethnic residential segregation in a multiethnic metro. *Sociological Methods & Research, 34*(2), 173–203.

59. Wilson, T., Grossman, I., Alexander, M., et al. (2022). Methods for small area population forecasts: State-of-the-art and research needs. *Population Research and Policy Review, 41*(3), 865–898.

60. Zhang, J. (2004). A dynamic model of residential segregation. *Journal of Mathematical Sociology, 28*(3), 147–170.

61. Zhang, J. (2004). Residential segregation in an all-integrationist world. *Journal of Economic Behavior & Organization, 54*(4), 533–550.

62. Zhang, J. (2011). Tipping and residential segregation: a unified schelling model. *Journal of Regional Science, 51*(1), 167–193.

63. Zou, Y., Fonoberov, V. A., Fonoberova, M., et al. (2012). Model reduction for agent-based social simulation: coarse-graining a civil violence model. *Physical Review E, 85*(6), 066106.

Authors and Affiliations

Yuchao Chen[1] · Yunus A. Kinkhabwala[2] · Boris Barron[1,3] [ID] · Matthew Hall[3,4] · Tomás A. Arias[1] · Itai Cohen[1]

✉ Itai Cohen
Itai.cohen@cornell.edu
https://scholar.google.com/citations?user=lDF0ZOUAAAAJ

Yuchao Chen
changyu@mit.edu
https://scholar.google.com/citations?user=UTGcfQoAAAAJ

Yunus A. Kinkhabwala
y.kinkhabwala@psehealthyenergy.org
https://scholar.google.com/citations?user=tsNCiroAAAAJ

Boris Barron
bb667@cornell.edu
https://scholar.google.com/citations?user=4KhbFbcAAAAJ

Matthew Hall
mhall@cornell.edu
https://scholar.google.com/citations?user=swvOzHwAAAAJ

Tomás A. Arias
taa2@cornell.edu
https://scholar.google.com/citations?user=Qk11UKEAAAAJ

[1] Department of Physics, Cornell University, Ithaca, NY 14850, USA

[2] Department of Applied and Engineering Physics, Cornell University, Ithaca, NY 14850, USA

[3] Cornell Population Center, Cornell University, Ithaca, NY 14850, USA

[4] Policy Analysis and Management, Cornell University, Ithaca, NY 14850, USA