Discovery of simple pattern-forming mechanisms in the development of settlements through data-driven model identification

Bartosz Prokop and Lendert Gelens
Laboratory of Dynamics in Biological Systems, Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium

Peter F. Pelz and John Friesen
Chair of Fluid Systems
TU Darmstadt, Darmstadt, Germany

(Dated: November 14, 2022)

Abstract: The rapid increase of population and settlement structures in the Global South during recent decades motivates the development of suitable models to describe their formation and evolution. This is usually achieved by developing detailed and complex models. In this work, we suggest that the development of settlements can be described by simple pattern-forming mechanisms. We provide a theoretical motivation for this hypothesis and select suitable spatio-temporal datasets of three regions of the Global South. We then investigate the dynamics of selected population density datasets using the data-driven white-box approach SINDy and provide explicit models directly from measured data. Even though at present this approach does not detect simple pattern-forming mechanisms in settlement development, we find the data resolution to be the limiting factor through a sensitivity study of SINDy applied to pattern-formation processes. Overall, the study provides a theoretical framework for the analysis of large-scale geodetic/ecological systems, which poses a promising view on studying complex urban systems and motivates further improvements in optimization approaches and data collection.

I. INTRODUCTION

The fraction of population living in urban or settlement structures has grown exponentially over the last several decades, especially in the Global South [1, 2]. This trend poses one of the main challenges in our world [3] as the rising population in such structures is in need of vital infrastructure while [4] simultaneously affecting (mostly negative) climatic developments [5, 6]. Consequently, there is an urgent need to understand underlying processes of urbanization and anticipate the emergence of these structures.

We know from the field of urban studies that urbanization and development of settlement structures depends on several mechanisms based on repulsion and attraction [7, 8]. Such interactions can lead to three major settlement distributions (see Fig. 1) of which the existence has been confirmed in recent settlement pattern studies of different regions in the Global South, in which regular distributions are dominating [9–13].

To understand the emergence of these interactions, different modeling approaches were developed. For example, urban development can be modeled by agent- or cellular-automata-based approaches [14, 15] which include detailed interactions at the level of individuals. Despite their accuracy in specific cases, such models have several drawbacks. They lack generalization, require large and detailed datasets to be fit on and become computationally expensive. Therefore, we suggest the application of pattern-forming mechanisms which not only include the described interaction mechanisms, e.g. reaction-diffusion models [16], but are simple and directly interpretable models that despite their simplicity are able to produce highly complex patterns and dynamics [17].

Furthermore, the existence of regularly-patterned distributions in settlements, and in other spatial systems, is seen as an indicator for the existence of instability-driven pattern-forming mechanisms [18]. A similar concept of linking spatial distributions with specific driving mechanisms has been successfully applied in the field of plant and animal ecology for decades [19, 20]. In the context of urban structures, Pelz et al. [21] have developed a theoretical framework describing the formation of informal settlements (so-called slums) in the Global South. Further, this framework was extended to describe the morphogenesis of urban systems in the United States as a reaction-diffusion system in [22].

In this work we intend to expand these ideas to rural settlements and investigate if it is possible to find sim-
ple pattern-forming mechanisms describing their emergence. To this end, we apply the system-identification approach called "Sparse Identification of Nonlinear Dynamics" (SINDy) [24] on spatio-temporal data from satellite images of selected regions of the Global South.

The structure of the work is as follows. We first motivate why simple pattern-forming mechanisms are able to capture and describe key mechanisms for the emergence of settlement structures. We do this using the example of a reaction-diffusion model which is able to create Turing patterns. We link physical aspects of the model in the context of actual socioeconomic behavior, as was done by Pelz et al. [22] (see Section II A). Here, we additionally prove that the postulated behavior can be seen in actual population density patterns. Afterwards, we introduce the datasets used for this investigation and their selection based on spatial regularity (see Section II B and II D). Following this, we briefly introduce SINDy and its extension for partial differential equations (called PDE-FIND [25]) and how the Akaike Information Criterion (AIC) is applied for model evaluation (see Section II C). In Section III, we present and evaluate identified model equations. Additionally, we analyze the dependence of the SINDy identification method in low-data limits using the example of a simple pattern-forming mechanism. Finally, in Section IV we summarize and discuss our findings and answer the question to what extent simple mechanisms can be identified from spatio-temporal data sets of the Global South and what has to be done to fully answer this question.

II. METHODS AND DATA

A. Interpretation of pattern-forming mechanisms in the context of settlements

As we postulate that simple mechanisms are responsible for the emergence of settlement structures, we start by motivating an exemplary interpretation of a reaction-diffusion mechanism for the emergence of rural settlements following the example of Pelz et al. [22].

We divide our system into interacting agents: a population living in the rural settlements and a supply potential of agricultural space within a domain of size $A$. The population density at a spatial point $x$, $t$ is given by $u(x,t)$ and the corresponding supply potential by $v(x,t)$. The increase rates $N_u$ and $N_v$ have three complementary contributions: (i) birth or death of the population or the cultivation or sealing of agricultural space within a domain of size $A$, (ii) migration to and from other cities outside of $A$ leading to a decrease or increase in supply potential and (iii) migration to areas with higher supply potential over the boundary $C$ of $A$ where a settlement exists or is created. Here, contributions (i) and (ii) can be considered as long-distance effects, while (iii) is a short-distance effect.

If we formulate the balance equations for the respective agents, applying Gauss’ theorem and providing the suitable transformations, we can achieve the well-known structure of the reaction-diffusion model (see also Appendix B):

$$u_t = \Delta u + Rf(u,v)$$
$$v_t = D\Delta v + Rg(u,v),$$

with $D, R$ being the diffusion and reaction coefficients alongside the respective reaction terms $f, g$. After conducting a linear stability analysis, we see that the dynamical behavior of the system is described by the signs of the Jacobi matrix of the reaction terms evaluated at the steady state, which are in the case of an instability as follows,

$$J = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix} = \begin{pmatrix} + & + \\ - & - \end{pmatrix}.$$ (2)

Following the model formulation we can therefore interpret the entries of the Jacobian in the following way:

- $f_u|_0 > 0$, people attract other people: The population $u$ increases due to self-reproduction of $u$ in an environment of sufficient sustenance.
- $f_v|_0 > 0$, supply attracts people: The amount of available agricultural area $v$ attracts people, increasing the population $u$ of the settlement.
- $g_u|_0 < 0$, people inhibit supply: The higher the population $u$, the less agricultural area $v$ is available, especially due to the limitation by the agricultural needs of surrounding settlements. This is the case until the maximum amount of agricultural area is used and does not suffice to supply the population of a settlement, leading to a decrease of population $u$.
- $g_v|_0 < 0$, supply inhibits additional supply: Due to limited resources, the supply production decreases, when agricultural efficiency plateaus.

In this situation a feedback loop is established between the size of a settlement or population and the size of required agricultural space surrounding as settlement or supply potential of a settlement, resulting in an instability.

This means that when analyzing the linear stability including diffusion, a stable and equally dispersed population and supply potential densities become unstable when following condition is met (see Appendix C):

$$f_u|_0 > -\frac{g_v|_0}{D}.$$ (3)

This is the case when the attraction of people to $A$ dominates the inhibition of supply potential due to the emergence of settlements. If the domain size $A$ and diffusion coefficient $D$ of the system are sufficiently large, different settlements can emerge that constantly compete against
each other over the supply potential, eventually leading to a regular distribution of settlements.

In our case, we interpret \( u, v \) as space or area occupied by population or agriculturally used space, which are not infinite and complementary. When we assume for \( \hat{U} \) and \( \hat{V} \) to be the maximum concentrations that can exist in one spatial point \( x \), this results in a maximum concentration \( \hat{u} + \hat{v} \). Hence, \( u \) and \( v \) can be transformed into each other: \( u = (\hat{U} + \hat{V}) - v \). Through this, we can apply the reaction-diffusion approach to our satellite data, where \( u \) is the population density and \( v \) the lack of it. Furthermore, the reaction-diffusion equation in Eq. [4] can be rewritten into a one-component, two-dimensional equation,

\[
\begin{align*}
    u_t(x, y, t) &= \hat{D} \Delta u(x, y, t) + R\hat{f}(u(x, y, t)), \\
    \text{with} \quad &\hat{D} = \frac{D + 1}{2}, \\
    \text{and} \quad &\hat{f} = \frac{1}{2} [f(u(x, y, t)) + g(u(x, y, t))].
\end{align*}
\]

This form of equation will be the basis for the term library in the later SINDy analysis. Additionally, as there are other one-component and two-dimensional equations, e.g. the Swift-Hohenberg [26] or Allen-Cahn [27, 28] equations, which can create patterns driven by instability, we extend the library by derivatives up to the fourth order resulting in a library shown in Tab. 1.

Additionally to this theoretical motivation for using pattern-forming mechanisms in the development of settlements, we can also observe the described dynamical behavior of populations (here of population density patterns from Worldpop) as shown in Fig. 2. Here, we show behavior like local migration through attraction of bigger settlements (e.g. between 2003 and 2008), invasion or occupation of available agricultural space (e.g. between 2008 and 2013) and local migration induced by competition of settlements over available space leading to changes in settlement patterns (e.g. between 2013 and 2018).

**B. Data processing and selection**

We selected three representative regions of emerging countries in the Global South for our investigation – the Punjab Region in India, the Nile delta in Egypt and the Kano State region in Nigeria (see Table 1). The regions have been chosen as they lie in countries which can be considered representative of the Global South: all countries have had steady population as well as steady economic growth over the last 20 years [1].

Despite and because of the great cultural differences, all three societies are in transition from agricultural countries to more industrialized nations. Furthermore, these regions have been the subject of aforementioned studies on spatial distributions of settlements [9][11] and are characterized by regularity on different spatial scales. This is of high importance for the subsequent steps in which we select sub-regions (here called regions of interest (ROI)) that show a regular distribution.

For the selection of the ROIs, we use the Global Artificial Impervious Area (GAIA) [30, 31] dataset, as it combines the currently best time (34 yearly time points between 1984 and 2018) and spatial resolution (30 m in East-West and North-South spatial resolution at the equator per pixel; for a comprehensive overview of spatio-temporal satellite datasets, see [32]). The GAIA dataset depicts areas that have been built-up in the form of a

| TABLE 1. Terms included in the library for the one-component, two-dimensional equation sorted by combinations of \( u \) and its derivatives. |
|---|
| Terms |
| Combinations | 1, \( u, u^2 \), \( u^n \) |
| Derivatives | \( u_x, u_y, u_{xx}, u_{xy}, u_{yy}, u_{xxy}, u_{yy}, u_{y}, u_{xxx}, u_{xxy}, u_{yyy}, u_{xxy}, u_{yy} \) |

FIG. 2. Example of multiple effects found in spatio-temporal data sets of population densities (with \( u_{max} \) being the maximum population density in the excerpt) which are described by the exemplary interpretation of RD equations. Effects are shown here in excerpt of spatio-temporally resolved population density data set Worldpop [29] which is introduced in Section 1. (Here an excerpt from the Punjab region in India). The observable behaviors are internal migration into existing settlements, invasion of not occupied agricultural space or local migration triggered by competition over available agricultural space.
FIG. 3. Exemplary workflow for the data selection of region of interest: i) We first select a region of the Global South, which is known to depict regular settlement structures. ii) Data acquisition of built-up structures in the GAIA data set of the Punjab. iii) Analysis of the GAIA data set with ANN and sliding window set up. The contour plots are made for the Punjab with a sliding window size of 15 km × 15 km. The ANN varies between 0 (clustered distribution) to 2.14 (regular distribution) with 1 being a random distribution. Three regular regions are being selected. iv) Extraction of ROIs from the GAIA and WorldPop datasets in the example of ROI 2 of India (Density shown on a logarithmic scale).

In contrast to the ROI selection process, we refrained from using the GAIA data set for the model identification as there is a discrepancy considering the patterns that are created by simple pattern-forming mechanisms. Most mechanisms describe concentration distributions of different (chemical or other) species and do not create discrete patterns. Therefore, we use a temporal data set of spatial population distribution provided by WorldPop. Currently this data set is the best available population density spatio-temporal data set with a spatial resolution of 100 m in East-West and North-South direction at the equator over 20 yearly time points.

To select the ROIs, we first performed an analysis of the ANN index in all three regions over time and with different sliding window sizes. For the ANN calculation we used the yearly data points of GAIA between 2000 and 2018 which overlap the WorldPop data set. Here, we select window sizes between 5 km over 5 km and 50 km over 50 km in region excerpts extending over 200 km in North-South and West-East direction. Second, following this method we create contour plots for all regions, window sizes and four selected years, which can be found in our Gitlab repository (https://gitlab.kuleuven.be/gelenslab/publications/settlements.git).

Resulting from this, we set the window size to 15 km and select three ROIs per region that show the highest ANN, meaning a regular settlement distribution (see Fig. III B iii). Lastly, with these selections, the ROIs have been extracted from the WorldPop data set for yearly data points between 2000 and 2020 (see Fig. III B iv).

The reason for this sequence of steps in order to select ROIs and moving from the built-up to the concentration distribution data lies in the data structure and the ANN analysis. The ANN index is calculated by evaluation of point distributions, where each point in the region represents an individual settlement. In order to generate these points individual settlements have to be distinguished, which is easily possible with a discrete signal but not with a continuous population distribution as provided by WorldPop. Indeed, WorldPop distributes population quantities in administrative regions and therefore the population concentration per pixel, which is never equal to 0 [29] and requires a sophisticated cluster algorithm in order to determine settlement centers which has

| Region           | Attributes                                                                 |
|------------------|-----------------------------------------------------------------------------|
| Punjab, India    | northwest India, border region with Pakistan, agricultural region called Granary of India |
| Nile delta, Egypt| north Egypt, densely populated, fast growing agricultural region             |
| Kano State, Nigeria | north Nigeria, border region to Niger, agricultural region in one of the fastest growing economies |
C. Model identification

The method we use in order to identify underlying mechanisms in the datasets is SINDy [24]. This approach has been increasingly applied in many fields, mainly in fluid mechanics. However, it has, to our knowledge, never been used for studying large scale, geo-sociological questions.

The main idea behind this method, is the assumption that dynamic systems can be described through either ordinary, or in our case partial differential equations (PDE- FIND) [25], with sparse structure in the following form:

\[ u_t = N(u(x,t), u_x, u_y, ..., x, \xi) \]  

(5)

The temporal change of \( u, u_t \), is a function of the variable \( u \) itself, its spatial derivatives and a set of coefficients \( \xi \). Differential equations of this form can be linearly combined:

\[ u_t = \xi_1 + \xi_2 u + \xi_3 u^2 + \xi_4 u_x + \xi_5 u_{xx} + ... \]  

(6)

This equation can be rewritten as a row vector containing all combinations and derivatives of the quantity, called the term library and a coefficient vector \( \xi \) containing all coefficients:

\[ u_t = \begin{pmatrix} 1 & u & u^2 & u_x & u_{xx} & \ldots \end{pmatrix} \xi. \]  

(7)

The values of each term in the library can be calculated from a single shot at a given point in time. If this system is extended to all available time points, a linear system of equations with the unknown parameter vector \( \xi \) and the term library matrix \( \Theta \) is formed:

\[ \begin{pmatrix} u_t \end{pmatrix} = \begin{pmatrix} 1 & u & u^2 & u_x & u_{xx} & \ldots \end{pmatrix} \begin{pmatrix} \xi \end{pmatrix} = \Theta \xi \]  

(8)

This system poses an over-determined optimization problem for values of \( \xi \) and can be solved using regression algorithms (for more detailed information on regression algorithms for SINDy, see [36]). In contrast to the original work on PDE-FIND [25], we apply a sparsity promoting algorithm with the SR3 algorithm developed by [37];

\[ \min_{\xi, w} \frac{1}{2} \| u_t - \Theta \xi \|^2 + \lambda \| w \|_1 + \frac{\alpha}{2} \| w - \xi \|^2 \]  

(9)

with \( \lambda = \frac{l^2}{2\alpha} \)

Here, two hyper-parameters of the optimization have to be set: the threshold \( l \) and the parameter of the optimization \( \alpha \) and the resulting penalizing parameter \( \lambda \) of the regularization.

D. Model selection with the Akaike Information Criterion

Following the model identification, we analyze the discovered models with the Akaike Information Criterion (AIC). The AIC is a measure of parsimony [38]. It compares the goodness of fit of a given model to the ground truth and weighs it with the model’s complexity aiming on maximizing the information provided by the simplest-as-possible model. For our analysis, we apply the corrected formulation for finite sample sizes of the AIC (AICc) proposed by Mangan et al. [39]:

\[ \text{AIC}_c = \text{AIC} + \frac{2(k + 1)(k + 2)}{m - k - 2} \]  

(10)

The AIC is described by the likelihood function of average error over time and space \( \epsilon \) as follows:

\[ \text{AIC} = m \ln(\epsilon/m) + 2k \]  

with \( \epsilon = \frac{\sum_{i=1}^{m} y_i - N(x_i, \xi)}{m} \)  

and \( m = m_t + m_s \)

The AIC depends on the amount of observations \( m \) (time points \( m_t \) and size of region \( m_s = XY \)) and the number \( k \) of terms describing the complexity of an identified model. With this we can determine the most parsimonious model among all models and study its properties. In order to compare the models, we further normalize the AIC by the minimal value of the respective analysis AIC_{min}. Here, following Mangan et al. [39] a model that has an AIC_c − AIC_{min} < 2 has strong support and with an AIC_c − AIC_{min} < 8 weak support for being the correct model. Hereafter, we always refer to the corrected AIC_c when the AIC is mentioned.

E. Data and Code Availability

All calculations, simulations and graphs are done in Python. For SINDy, we use the package PySINDy [40, 41] and for simulations, we developed a simple forward-Euler solver. All algorithms are available in our Gitlab repository [https://gitlab.kuleuven.be/gelenslab/publications/settlements.git].

III. RESULTS

We scan sets of thresholds from \( l = 10^{-6} \) to \( l = 10^2 \) and optimization hyper-parameter \( \alpha = 10^{-3} \) to \( \alpha = 10^3 \) for our investigation. The SR3 algorithm is applied with a tolerance of \( 10^{-2} \) and 200 iterations (For details see [37]).

From the parameter scan, we determine for which combinations of \( l \) and \( \alpha \) the identified models provide an
FIG. 4. Parameter sweep of the all ROIs and regions for set of threshold \( l \) and optimization parameter \( \alpha \). (i) The analysis of the ROIs shows different combinations of parameters where the AIC falls beneath 2 for the respective ROI. As we search for a regionally valid solution, we select only unique solutions at combinations of \( l, \alpha \) where the AIC < 2 for all ROIs overlap. The selected unique equations and their respective optimization parameter sets where they were found first are shown. Red border markers depict the best identified models of each ROI. We see that the AIC provides the best equations with the lowest error and complexity (except for ROI 2, from which we show later through analysis of contributions (Section III A) that only the production terms are significant) (ii) The respective error and complexity of the found equations are depicted. The red bordered markers show the best model for each ROI in each region following the AIC. (iii) The identified coefficients \( \xi \) (as in Eq. (8)) are shown for each equation. All identified models contain most of the derivatives (diffusive terms) with large coefficient values. If found the coefficient values of the reaction terms are multiple magnitudes smaller than of the diffusive terms.
AIC < 2. In order to pre-select possible solutions, we only evaluate equations with this AIC at overlapping parameter pairs of $l$ and $\alpha$. The reason for this is, that if the dynamics in the whole region follow the same rules (or dynamical behavior), all best identified models should have the same mechanistic form. Therefore, in Fig. 4 we show for which sets of $(l, \alpha)$ the low-AIC regions of each ROI overlap (in yellow).

Following this, we selected unique model equations at their lowest parameter values for $l$ and $\alpha$ respectively. The parameter combinations for unique solutions of each ROIs are shown in Fig. 4i. These unique solutions are further compared with the use of the AIC, where we also depict the error and the complexity of the found equations in Fig. 4ii. The optimal selected equations are highlighted by markers with red borders and the values of coefficients are shown in Fig. 4iii.

A. The discovered models are dominated by production

The results of the analysis show that when applying SINDy, the identified models are not sparse and do not represent any known simple pattern-forming mechanism. The assigned coefficients vary in multiple order of magnitude, between $10^{-3}$ for some production terms up to $10^6$ for some higher order derivatives of the diffusive terms.

This selection is logical, as the values of higher order derivatives are small compared to production terms and therefore need a high coefficient to provide significant contribution to the overall dynamics. However, when trying to determine a sparse model in a low-data limit (low temporal and spatial resolution, small size of temporal samples), the optimization algorithm overestimates the importance of some higher order derivatives. Therefore, we need to investigate the actual contribution of each term to the overall dynamics of a model. For this, we calculated the contribution as the mean value of each term at a set time point in the calculation (we arbitrarily chose $t = 10$), which is multiplied with the respective coefficient,

$$c_i = \left| \frac{\sum_{x=0}^{X} \sum_{y=0}^{Y} \theta_{ij}(x, y, t)}{XY} \xi_j \right| \quad \text{with} \quad t, i = 10 \quad (12)$$

The contribution of each term and for all equations is shown in Fig. 5i. From this we see that even though the coefficients of derivative terms were multiple orders of magnitude larger than the production terms, their contribution is negligible and lies between order of $10^{-17}$ and $10^{-21}$. (ii) The domination of production terms can be also seen in the comparison between the found models and the actual data set. This leads to an increase of concentration in the whole region, but no pattern formation. Despite being trained on the data set, the models are not able to reproduce the training data (Here ROI 2, India).
vation is general: all models do not capture the actual dynamics (the results for the respective ROIs can be found in our repository).

C. Spatio-temporal resolution strongly influences model recovery

As can be seen from the above analysis, the identification of pattern-forming mechanisms in satellite imagery data has not lead to the identification of a clear mechanism or a model known from literature. We suspect that this lack of identification results from the low spatio-temporal resolution of our data. From other works, it is known, that the model derivation with SINDy depends on the amount of available temporal points and the size of the time step $dt [37, 42]$. The time step of WorldPop is limited to one snapshot per year, which can lead to a faulty parameter estimation. We derive, that this has also to be true for spatial points, even though to our knowledge the limits of SINDy in low-data limits, meaning low spatial and temporal resolution have not been investigated in depth. We therefore decided to conduct a brief study on this topic using an equation of the form suggested in Eq. (4).

We chose the aforementioned Allen-Cahn (AC) equation [27, 35], which is the non-mass conserved version of the Cahn-Hilliard equation [28], a simple and well-studied one-component model for pattern formation,

$$u_t(x, y, t) = \alpha \Delta u + \beta u + \gamma u^2 - u^3 \tag{13}$$

It can be interpreted in the notation of Eq. (4) as follows:

$$\bar{D} \Delta u(x, y, t) = \alpha \Delta u,$$

$$R \bar{f}(u(x, y, t)) = \beta u + \gamma u^2 - u^3. \tag{14}$$

With this set of parameters and the initial condition $u_{init}(x, y, t = 0) \sim N(0, 0.01)$, Eq. (13) creates coarsening labyrinth patterns shown in Fig. 6i/ii. For the evaluation with SINDy, we rewrite the equation as a combination of terms from Tab. 1

$$u_t = \xi_1 u + \xi_2 u^2 + \xi_3 u^3 + \xi_6 u_{xx} + \xi_7 u_{yy} \tag{15}$$

and the other parameters are equal to 0. The evaluation of the dataset with a spatial resolution of $dx, dy = 0.39$ $(256 \times 256)$ and temporal resolution of 1250 frames or $dt = 0.08$ leads to the following discovered equation:

$$u_t = \xi_1 u + \xi_2 u^2 + \xi_3 u^3 + \xi_6 u_{xx} + \xi_7 u_{yy} \tag{16}$$

The recovered equation has the exact form as the initial Allen-Cahn equation in Eq. (13), with a maximum error of 3% in the coefficients.

Next, we subsampled the data temporally and spatially, in order to identify the limits of the SINDy approach. We decreased the amount of temporal samples from 1250 ($dt = 0.08$) in 26 steps to 10 samples ($dt = 10$). Simultaneously, the amount of spatial samples was reduced from 65536 spatial points (256 $\times$ 256 points) in 18 steps to 100 spatial points (656 $\times$ 656 points) in 18 steps. The recovered equation has the exact form as the initial Allen-Cahn equation in Eq. (13), with a maximum error of 3% in the coefficients.

The recovered equation has the exact form as the initial Allen-Cahn equation in Eq. (13), with a maximum error of 3% in the coefficients.

Next, we subsampled the data temporally and spatially, in order to identify the limits of the SINDy approach. We decreased the amount of temporal samples from 1250 ($dt = 0.08$) in 26 steps to 10 samples ($dt = 10$). Simultaneously, the amount of spatial samples was reduced from 65536 spatial points (256 $\times$ 256 points) in 18 steps to 100 spatial points (656 $\times$ 656 points) in 18 steps. The recovered equation has the exact form as the initial Allen-Cahn equation in Eq. (13), with a maximum error of 3% in the coefficients.

Next, we subsampled the data temporally and spatially, in order to identify the limits of the SINDy approach. We decreased the amount of temporal samples from 1250 ($dt = 0.08$) in 26 steps to 10 samples ($dt = 10$). Simultaneously, the amount of spatial samples was reduced from 65536 spatial points (256 $\times$ 256 points) in 18 steps to 100 spatial points (656 $\times$ 656 points) in 18 steps. The recovered equation has the exact form as the initial Allen-Cahn equation in Eq. (13), with a maximum error of 3% in the coefficients.
steps to 196 spatial points (14 × 14 points). We then compared the identified model to the original dynamical model, determining if the actual mechanism (coefficients are correct) or the mechanistic form of the model (coefficients are not, but terms are correct) was detected.

The results of the sensitivity study can be seen in Fig. 6. As expected, the SINDy algorithm is highly sensitive to spatial and temporal resolutions. Only for high spatial and temporal resolutions of above 45796 spatial points (214 × 214) and \( dt = 2 \) (50 frames) we are able to correctly recover the equation of the underlying dynamics. For lower resolution, we are still able to recover the mechanistic form of the equations. The results of our sensitivity study fit well with the structure of models identified with the SINDy analysis. When decreasing the resolution, the optimization algorithm includes and overestimates higher order derivatives.

Such models are not able to capture the dynamics of a system accurately, which we show in Fig. 6(i). Here, we simulated three cases from the sensitivity analysis (see Fig. 6). For the row labeled "O", we show the original time series of the system. In row "1", we show the case of high-temporal/low-spatial resolution, while in row "2" we explore the case of high-temporal/low-spatial resolution. Finally, in row "3", we show the case of insufficient resolution, and note that this resolution is still better than the satellite dataset we used. For cases 1 and 2, where the actual mechanistic form is still correct, we see that the overall dynamics of the system can be still recovered (when compared to the provided dataset). The reduction of spatial resolution leads to coefficients smaller than those in the original equation, which slows down the dynamics while still preserving the behavior. The reduction of temporal points has no significant influence on the recovery of the dynamics, as the decrease of coefficient values is not as stark as for the reduction of spatial resolution. For case 3, we find that when evaluating the equation on the initial time interval of \( t = [0, 100] \), the identification of the dynamical system is severely obstructed, while also recovering a non-sparse solution with the SINDy algorithm. This becomes even more challenging when considering noisy and even more sparsely sampled experimental data.

D. Successful recovery requires sufficient observation time capturing dynamical changes

Aside of the influence of spatial and temporal resolution, we also explore the influence of the observation time on the identification of underlying dynamics. This is motivated by the fact that we only observe relatively slow changes of settlement structures over 20 years using the WorldPop dataset. To examine how the observation time influences model discovery, we again turn to the Allen-Cahn dynamics for which we analyzed the role of temporal and spatial discretization in model identification, see Fig. 6. Here, we fix the time discretization \( (dt = 0.16) \), which allowed for a successful identification of the equation in Fig. 6, but progressively reduced the observation time (or equivalently, the fraction of time frames included) of the time series used in the SINDy optimization. We start with all time points \( t = [1, 100] \) (626 images or fraction of 1) until \( t = [99, 100] \) (6 images or fraction of ca. 0.01).

The results of this analysis are shown in Fig. 7. When reducing the observation time, thus capturing less dynamical changes, below a fraction of images of \( \approx 75\% \) it is no longer possible to identify the correct underlying dynamics anymore. In the case of our settlement dataset, it is thus plausible that the observed changes in population density are inadequate for proper model identification due to the low temporal resolution and the short observation time compared to the relevant time scales over which settlements develop.

IV. DISCUSSION AND CONCLUSION

The goal of this work was to not only theoretically describe the possible role of simple pattern-forming mechanisms in the development of urban structures, as has been done before by [22, 23], but to provide an as unbiased as possible approach to identify such models directly from data. In order to do this, we extended the ideas of [22], motivating and providing a new theoretical point of view on pattern-forming mechanisms in rural, agriculturally dominated settlement structures. We argued that together with features of regularity (as suggested by [9, 18, 19]), pattern-forming mechanisms should and can be considered responsible for the emergence of settlement structures. This extension can be a starting point in crit-
ically evaluating urban modeling approaches that strive for more complexity over generalization.

Furthermore, we analyzed the occurrence of regularity in settlement structures of selected regions of the Global South (i.e., India, Egypt, and Nigeria). Using these data, we selected regions of interest (ROIs) in the spatio-temporal dataset WorldPop of population density distributions.

We then introduced the SINDy method together with the AIC, allowing us to derive and investigate spatio-temporal models for the dynamics of population density patterns. We are, to our knowledge, the first to attempt model identification for large scale geodetic dynamical system directly from satellite data. Even though data-driven white and gray box methods like SINDy have become very popular in recent years, there are not widely applied for model identification from actual experimental data in general, especially not on geodetic or ecological systems.

We decided to evaluate the role of each term by calculating its relative contribution to the dynamics, which neither sparse nor represented known pattern-forming mechanisms from literature. The assigned coefficients differed in multiple orders of magnitude between production terms ($10^{-3}$) and diffusion terms ($10^9$).

Here two aspects play a major role for our analysis. First, we recovered models that were not sparse and contained terms that despite having large coefficients did not significantly contribute to the dynamics. As a result, we have to change the target of optimization from solely evaluating the coefficients to targeting the actual contribution of terms, which e.g. is estimated by [43]. Additionally, the configuration of our SINDy approach did not include any time or space dependency of parameters, as suggested by [44], which could pose a significant disadvantage to capture important dynamical behavior of settlement systems in the Global South. We decided to evaluate the role of each term by calculating its relative contribution to the dynamics, which showed that the dynamics are dominated by the production terms and cannot be understood as pattern-forming mechanisms, while also not being able to satisfactorily capture the dynamics of the training data.

The second aspect considers the quality of data, which we used for our investigation. We suggested that the used dataset had too low spatial and temporal resolution, and accordingly carried out a sensitivity study of SINDy towards low-data limits. We showed that SINDy (here PDE-FIND) is sensitive to spatial and temporal resolution, and the currently available population density datasets such as WorldPop do not provide sufficient resolution to conclusively answer our research question. Here, we considered to improve the quality by interpolating the data to more spatial and temporal points, but quickly discarded it, as applying a regression-based algorithm to this artificial data would uncover the introduced interpolation, but not the actual dynamics. Additionally, we found that the observation time of the dynamical system has a strong influence on the recovery of underlying dynamics using the SINDy algorithm as the time series need to be sufficiently long compared to the relevant temporal time scales of the system dynamics. Hence, we need to closely follow the rapid improvements in data acquisition with satellite imagery, which would provide us with sufficiently good data, in which case our work provides a ready-to-use framework. Additionally, the structure of the data should be adjusted. As mentioned before, the current WorldPop dataset does not allow for uninhabited areas with a population density of 0. Here, WorldPop itself is currently working on providing an improved dataset, where population densities and built-up areas are mapped. At the moment of publication this dataset only contains a single time point but, when extended, it will provide new opportunities to study our question.

In conclusion, we set a starting point towards the evaluation and identification of the role of simple pattern-forming mechanisms in the development of settlement structures. So far, the efforts were unfruitful to provide model equations in the form of such mechanisms. However, we provided a theoretical motivation and created a framework to answer this research question conclusively when higher resolution data becomes available.

ACKNOWLEDGMENTS

The work of the author J.F. is funded by the LOEWE Program of Hesse State Ministry for Higher Education, Research and the Arts within the project "Uniform detection and modeling of slums to determine infrastructure needs".

B.P. wrote the manuscript and code, conducted the analysis, developed the interpretation and created the figures, P.F.P. contributed to conceptualization, L.G. contributed to the manuscript, conceptualization and supervision and J.F. contributed to the manuscript, conceptualization and supervision.

Appendix A: Region Information

Here we attach the geographical data of the regions used to demonstrate the workflow of our method, see Table I. The coordinates are given in decimal degrees in reference system WGS84.
The long-distance effects are a product of the reaction–diffusion eigenvalue problem with the eigenvalue stability analysis around the linearized state of Eq. (1). As done in [22] and following [35] we perform a linear stability analysis with the perturbation ansatz $\delta \hat{u} = \mathcal{R} [\delta \hat{u} \exp(\sigma t + ikx)]$ or vice-versa with $v$, we derive an eigenvalue problem with the eigenvalue $\sigma$, the Kronecker delta $\delta$, $\mathbf{u} = (u, v)$, the Jacobian $\mathbf{J}(f, g)$ and $\mathbf{D} = 0$:

$$ (\sigma \delta - \mathbf{J}(f, g)) \delta \hat{u} = 0, \quad (C1) $$

$$ \sigma^2 - \mathbf{J}(f, g) \mathbf{I} \sigma + \det(\mathbf{J}(f, g)) = 0. \quad (C2) $$

Solving the eigenvalue problem results in two conditions for the Jacobian $\mathbf{J}(f, g)$,

$$ \mathbf{J} = \begin{pmatrix} f_u |_0 & f_v |_0 \\ g_u |_0 & g_v |_0 \end{pmatrix}, \quad (C3) $$

that lead to instability,

$$ f_u |_0 + g_v |_0 < 0, \quad (C4) $$

$$ \det(\mathbf{J}(f, g)) = g_u |_0 f_u |_0 - g_u |_0 f_v |_0 > 0. \quad (C5) $$

As described in [22] the only reasonable formulation of the Jacobian is:

$$ \mathbf{J} = \begin{pmatrix} + & + \\ - & - \end{pmatrix}. \quad (C6) $$

Other forms where the column-wise signs are the same result in concentrations spatially in phase and the form with row-wise same signs only the shown can be suitably used as shown in Section II.A

As Turing patterns can arise due to diffusion, we as well study the short-distance effects. We can reformulate diffusion as a product of specific energy $k_B T$, with the Boltzmann constant $k_B$ and temperature $T$, and the mobility $\mu$. At constant $T$, the ratio $D = \mu_v / \mu_u$ results in,

$$ \mathbf{B} := \mathbf{J}(f, g) - D \mathbf{k} \quad \text{with} \quad \mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & D \end{pmatrix} \quad (C7) $$

and allows to rewrite the eigenvalue problem in Eq. (C1) to:

$$ (\sigma \delta - \mathbf{B}(f, g)) \delta \hat{u} = 0. \quad (C8) $$

This results again in two conditions for the Jacobian,

$$ (f_u |_0 + g_v |_0 - k^2(1 + D) < 0, \quad (C9) $$

$$ \det(\mathbf{B}) = (f_u |_0 - k^2)(g_v |_0 - k^2) - g_u |_0 f_v |_0 > 0. \quad (C10) $$

Turing instability is achieved when the condition Eq. (C10) is violated, resulting in the necessary condition for the diffusion induced instability:

$$ D f_u |_0 + g_v |_0 > 0 \quad \rightarrow \quad f_u |_0 > -\frac{g_v |_0}{D} \quad (C11) $$

| Region | West | South | East | North |
|--------|------|-------|------|-------|
| India  | 75.3855 | 28.8265 | 77.4804 | 30.6380 |
| ROI 1  | 76.0086 | 29.4155 | 75.8528 | 29.5319 |
| ROI 2  | 76.7096 | 29.7066 | 76.5538 | 29.8230 |
| ROI 3  | 77.2547 | 30.2305 | 77.0990 | 30.3469 |
| Egypt  | 29.8650 | 29.9671 | 32.1010 | 31.8803 |
| ROI 1  | 30.3377 | 30.7896 | 30.1804 | 30.9044 |
| ROI 2  | 31.1260 | 30.6174 | 30.9686 | 30.7322 |
| ROI 3  | 31.5200 | 31.2488 | 31.3627 | 31.3636 |
| Nigeria | 7.3774  | 11.1881 | 9.3336  | 10.1556 |
| ROI 1  | 8.2020  | 11.9172 | 8.0646  | 12.0492 |
| ROI 2  | 8.2708  | 12.5114 | 8.1333  | 12.6434 |
| ROI 3  | 8.9580  | 12.1152 | 8.8205  | 12.2473 |

**Appendix B: Formulation of reaction-diffusion equations**

With the definitions from Section II.A we formulate balance equations for the respective agents $u, v$:

$$ \dot{N}_u = \frac{\partial}{\partial t} \int_A u \, dA = \int_A \dot{U} R f(u, v) - \oint_C \mathbf{J}_u \cdot \mathbf{n} \, dC, $$

$$ \dot{N}_v = \frac{\partial}{\partial t} \int_A v \, dA = \int_A \dot{V} R g(u, v) - \oint_C \mathbf{J}_v \cdot \mathbf{n} \, dC. \quad (B1) $$

The long-distance effect are a product of the reaction terms $f(u, v)$ or $g(u, v)$ and the reaction rates $\dot{U} R, \dot{V} R$. Here, $u := u / \bar{U}$ and $v := v / \bar{V}$ are dimensionless by division with reference desities $\bar{U}, \bar{V}$.

Similarly to [22], the short-distance effects are also driven by a density gradient which can be modeled with Fick’s first law. By applying Gauss’ theorem, we get the two reaction diffusion equations:

$$ \dot{u} = \dot{U} R f(u, v) + D_u \Delta u, \quad v = \dot{V} R g(u, v) + D_v \Delta v. \quad (B2) $$

With the additional dimensionless transformations $t := Rt$, $\mathbf{x} = \mathbf{x} \sqrt{R / D_u}$, $D := D_v / D_u$ we derive the dimensionless standard form of reaction-diffusion equations:

$$ \dot{u} = \Delta u + R f(u, v) \quad v = \Delta v + R g(u, v) \quad (B3) $$

**Appendix C: Linear stability analysis**

As done in [22] and following [35] we perform a linear stability analysis around the linearized state of Eq. (1) with $u = U + \delta u, v = V + \delta v$ with the homogenous solutions $U, V$. With the the perturbation ansatz $\delta u = \mathcal{R} [\delta \hat{u} \exp(\sigma t + ikx)]$ or vice-versa with $v$, we derive an eigenvalue problem with the eigenvalue $\sigma$, the Kronecker delta $\delta$, $\mathbf{u} = (u, v)$, the Jacobian $\mathbf{J}(f, g)$ and $\mathbf{D} = 0$:

$$ (\sigma \delta - \mathbf{J}(f, g)) \delta \hat{u} = 0, \quad (C1) $$

$$ \sigma^2 - \mathbf{J}(f, g) \mathbf{I} \sigma + \det(\mathbf{J}(f, g)) = 0. \quad (C2) $$

TABLE III. Coordinates of the regions of interest.

| ROI | West | South | East | North |
|-----|------|-------|------|-------|
| 1   | 8.2020 | 11.9172 | 8.0646 | 12.0492 |
| 2   | 8.2708 | 12.5114 | 8.1333 | 12.6434 |
| 3   | 8.9580 | 12.1152 | 8.8205 | 12.2473 |
[1] UN. World population prospects - average annual rate of population change (percentage) (2019).
[2] UN. Population facts: Policies on spatial distribution and urbanization have broad impacts on sustainable development (2020).
[3] F. Retief, A. Bond, J. Pope, A. Morrison-Saunders, and N. King. Global megatrends and their implications for environmental assessment practice. Environmental Impact Assessment Review 61, 52 (2016).
[4] E. A. Adams, J. Stoler, and Y. Adams, Water insecurity and urban poverty in the global south: Implications for health and human biology, American journal of human biology: the official journal of the Human Biology Council 32, e23368 (2020).
[5] H. Nagendra, X. Bai, E. S. Brondizio, and S. Lwasa. The urban south and the predicament of global sustainability, Nature Sustainability 1, 341 (2018).
[6] S. Thacker, D. Adshead, M. Fay, S. Hallegatte, M. Harvey, H. Meller, O. Regan, J. Rozenberg, G. Watkins, and J. W. Hall. Infrastructure for sustainable development, Nature Sustainability 2, 324 (2019).
[7] J. C. Hudson, A location theory for rural settlement. Annals of the Association of American Geographers 59, 365 (1969).
[8] W. Christaller, Die zentralen Orte in Süddeutschland (German), Central places in Southern Germany (English), 3rd ed. (Wissenschaftliche Buchgesellschaft, 1933).
[9] K. Henn, J. Friesen, J. Hartig, and P. F. Pelz, Spatial analysis of settlement structures to identify pattern formation mechanisms in inter-urban systems, ISPRS International Journal of Geo-Information 9, 541 (2020).
[10] A. A. AbouKorin, Spatial analysis of the urban system in the nile valley of egypt, Ain Shams Engineering Journal 9, 1819 (2018).
[11] B. Prokop and J. Friesen, Spatio-temporal interurban regularities in the global south, Preprints 10.20944/preprints202104.0752.v1 (2021).
[12] R. Yang, Q. Xu, and H. Long, Spatial distribution characteristics and optimized reconstruction analysis of china’s rural settlements during the process of rapid urbanization, Journal of Rural Studies 47, 413 (2016).
[13] J. Friesen, H. Taubenböck, M. Wurm, and P. F. Pelz, The similar size of slums, Habitat International 73, 79 (2018).
[14] C. Losiri, M. Nagai, S. Ninsawat, and R. Shrestha, Modeling urban expansion in bangkok metropolitan region using demographic-economic data through cellular automata-markov chain and multi-layer perceptron-markov chain models, Sustainability 8, 686 (2016).
[15] M. Batty and R. Milton. A new framework for very large-scale urban modelling, Urban Studies 58, 3071 (2021).
[16] A. M. Turing, The chemical basis of morphogenesis. 1953, Bulletin of Mathematical Biology 52, 153 (1990).
[17] R. M. May, Simple mathematical models with very complicated dynamics, The Theory of Chaotic Attractors , 85 (2004).
[18] R. M. Pringle and C. E. Tarnita, Spatial self-organization of ecosystems: Integrating multiple mechanisms of regular-pattern formation, Annual review of entomology 62, 350 (2017).
[19] C. E. Tarnita, J. A. Bonachela, E. Sheffer, J. A. Gutzton, T. C. Coverdale, R. A. Long, and R. M. Pringle, A theoretical foundation for multi-scale regular vegetation patterns, Nature 541, 398 (2017).
[20] C. Grohmann, J. Oldeland, D. Stoyan, and K. E. Linsenmair, Multi-scale pattern analysis of a mound-building termite species, Insectes Sociaux 57, 477 (2010).
[21] G. Theraulaz, E. Bonabeau, S. C. Nicolas, R. V. Solé, V. Fourcassié, S. Blanco, R. Fournier, J.-L. Joly, P. Fernández, A. Grimal, P. Dalle, and J.-L. Deneubourg, Spatial patterns in ant colonies, Proceedings of the National Academy of Sciences of the United States of America 99, 9645 (2002).
[22] P. F. Pelz, J. Friesen, and J. Hartig. Similar size of slums caused by a turing instability of migration behavior, Physical Review E 99, 022302 (2019).
[23] J. Friesen, R. Tessmann, and P. F. Pelz. Reaction-diffusion model describing the morphogenesis of urban systems in the us, Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management - GISTAM (2019).
[24] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences 113, 3932 (2016).
[25] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations, Science advances 3, e1602614 (2017).
[26] J. Swift and P. C. Hohenberg. Hydrodynamic fluctuations at the convective instability, Physical Review A 15 (1977).
[27] S. M. Allen and J. W. Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening, Acta Metallurgica 27, 1085 (1979).
[28] J. W. Cahn and J. E. Hilliard. Free energy of a nonuniform system. i. interfacial free energy, The Journal of Chemical Physics 28, 258 (1958).
[29] Worldpop, Worldpop - population density (2019).
[30] P. Gong, X. Li, J. Wang, Y. Bai, B. Chen, T. Hu, X. Liu, B. Xu, J. Yang, W. Zhang, and Y. Zhou. Annual maps of global artificial impervious area (gaia) between 1985 and 2018, Remote Sensing of Environment 236, 111510 (2020).
[31] X. Liu, Y. Huang, X. Xu, X. Li, X. Li, P. Ciais, P. Lin, K. Gong, A. D. Ziegler, A. Chen, P. Gong, J. Chen, G. Hu, Y. Chen, S. Wang, Q. Wu, K. Huang, L. Estes, and Z. Zeng. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015, Nature Sustainability 3, 564 (2020).
[32] D. R. Thomson, D. A. Rhoda, A. J. Tatem, and M. C. Castro. Gridded population survey sampling: a systematic scoping review of the field and strategic research agenda, International Journal of Health Geographics 19, 34 (2020).
[33] P. J. Clark and F. C. Evans. Distance to nearest neighbor as a measure of spatial relationships in populations, Ecology (1954).
[34] J. D. Murray, Mathematical biology. I Spatial models and biomedical applications (Springer, 2003) p. 811.
[35] M. C. Cross and P. C. Hohenberg. Pattern formation outside of equilibrium, Reviews of Modern Physics 65, 854 (1993).
[36] K. Champion, P. Zheng, A. Y. Aravkin, S. L. Brunton, and J. N. Kutz, A unified sparse optimization framework to learn parsimonious physics-informed models from data, *IEEE Access* **8**, 169259 (2020).

[37] P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin, A unified framework for sparse relaxed regularized regression: Sr3, *IEEE Access* **7**, 1404 (2019).

[38] H. Akaike, Information theory and an extension of the maximum likelihood principle, In proceedings of 2nd International Symposium on Information Theory, 267 (1973).

[39] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor, Model selection for dynamical systems via sparse regression and information criteria, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**, 10.1098/rspa.2017.0009 (2017).

[40] A. Kaptanoglu, B. de Silva, U. Fasel, K. Kaheman, A. Goldschmidt, J. Callaham, C. Delahunt, Z. Nicolaou, K. Champion, J.-C. Loiseau, J. Kutz, and S. Brunton, Pysindy: A comprehensive python package for robust sparse system identification, *Journal of Open Source Software* **7**, 3994 (2022).

[41] B. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. Kutz, and S. Brunton, Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data, *Journal of Open Source Software* **5**, 2104 (2020).

[42] S. Thaler, L. Pachler, and N. A. Adams, Sparse identification of truncation errors, *Journal of Computational Physics* **397**, 108851 (2019).

[43] G. T. Naozuka, H. L. Rocha, R. S. Silva, and R. C. Almeida, Sindy-sa framework: enhancing nonlinear system identification with sensitivity analysis, *Nonlinear Dynamics*, 1 (2022).

[44] S. Rudy, A. Alla, S. L. Brunton, and J. N. Kutz, Data-driven identification of parametric partial differential equations, *SIAM Journal on Applied Dynamical Systems* **18**, 643 (2019).