Computational and Applied Mathematics (2024) 43:420
https://doi.org/10.1007/s40314-024-02927-9

Building uncertainty-aware mathematical models based on evidence from datasets using grammatical evolution optimization techniques: the case of the obesity dynamics

Daniel Parra[1] · J. Ignacio Hidalgo[1] · José-Manuel Velasco[1] · Rafael-Jacinto Villanueva[2]

Received: 15 February 2024 / Revised: 29 August 2024 / Accepted: 2 September 2024

Abstract
This paper introduces a methodology to build mathematical models based on evidence and data sets, considering data and model uncertainty. We study the evolution of obesity in the population, being obesity a consequence of the transmission of unhealthy lifestyle habits and behavioral patterns influenced by social networks (family, friends, peers, etc.). We propose a three-step methodology. First, we create a synthetic data set based on a previous model with real data. Then, we search for dynamic models based on difference equations that best fit the dynamics described by the dataset and their uncertainty (uncertainty-aware). To do this, we use a dynamic structured grammatical evolution algorithm (an algorithm that builds possible models) on which we have defined a grammar (set of possible expressions that can be part of the model). The definition of appropriate grammar is crucial because it allows us to build models that do not contradict the knowledge of the phenomenon studied. However, the data may suggest introducing new terms that indicate the influence of unknown factors. Finally, from among all the models obtained, we will algorithmically search for a selection of them that best describes the uncertainty of the data. This methodology can be applied to various scenarios with available datasets and a limited understanding of the phenomenon. It aims to generate models that not only achieve precision but also incorporate terms that correspond to identifiable processes, which can be explained within the context of the study problem.

Keywords Datasets · Automatic model building · Dynamic models · Uncertainty quantification

Mathematics Subject Classification 37A50 · 37-11 · 92D30 · 68Q32 · 68W50

✉ Rafael-Jacinto Villanueva
rjvillan@upv.es

[1] Departamento de Arquitectura de Computadores, Universidad Complutense de Madrid, Madrid, Spain

[2] Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Valencia, Spain

Published online: 25 September 2024

1 Introduction

Obesity is a health condition linked to various health comorbidities, including diabetes, heart diseases, certain types of cancers, and more (Centers for Disease Control 2022; WHO Discussion Paper 2022). Several studies (Hill and Peters 1998; Christakis and Fowler 2007; Cohen-Cole and Fletcher 2008; Leahey et al. 2011) found that individuals were more likely to become obese if their friends, siblings, or spouses were obese. This social contagion aspect implies that obesity could be transmitted through social interactions and influence behavior patterns, making it a potential target for preventive measures. Therefore, it seems natural to model obesity as a socially transmitted disease and use the background that provides the mathematical epidemiology. Following this idea, several epidemiological-type mathematical models describing the transmission dynamics of obesity have appeared in the literature (Evangelista et al. 2004; Wadhera et al. 2016; Frerichs et al. 2013; Lozano-Ochoa et al. 2017; Hill et al. 2010; Santonja et al. 2010; Santonja and Shaikhet 2014).

Real-world data is often characterized by incompleteness, noise, and various sources of uncertainty, making it crucial to incorporate uncertainty into the modeling process to ensure robust and reliable predictions. Mathematical models serve as simplifications of real-world phenomena, with the objective of being able to provide knowledge about the problem under study. These simplifications may arise from the inability to consider relevant aspects without sufficient evidence or from disregarding less influential factors related to the phenomenon under study. These forms of uncertainty are referred to as epistemic or structural uncertainty. In addition, data itself can be subject to inaccuracies, high variability, and inherent uncertainty, which is known as aleatoric uncertainty (Smith 2013).

While models should be supported by evidence, there are instances where new terms may be introduced into the model that are not directly supported by the evidence but do not contradict it either. These additional terms can enhance the ability of the model to accurately explain the data and its associated uncertainty. In this way, we include the known and the unknown in the model. This idea will guide the developed technique in this paper.

However, it is not viable to continuously add or remove terms from the model without a defined strategy or to rely on trial and error to identify the most suitable options for explaining the data. It is essential to adopt a systematic approach based on algorithms that guide us in obtaining models that align with the evidence and effectively explain the data and its uncertainty.

This paper aims to build upon prior research by developing a mathematical model that explicitly incorporates uncertainty (uncertainty-aware) and, at the same time, integrates evidence from real-world datasets. To achieve this goal, we employ a three-step methodology:

- Firstly, we create a synthetic dataset derived from a previously established model, which serves as the basis for our analysis. We take as a base model the one in Santonja et al. (2012), where an analysis of the effect of public health campaigns on reducing excess weight using a mathematical model is performed, and data series in the Spanish region of Valencia are provided.
- Secondly, considering only the built synthetic dataset, multiple difference equation models are then generated using dynamic structured grammatical evolution (DSGE) (Lourenço et al. 2017, 2018), a powerful optimization technique that leverages evolutionary algorithms. This approach allows us to explore a broad range of potential mathematical models and their corresponding solutions by properly limiting the grammar to be compatible with the knowledge of the studied phenomenon.

- Finally, we employ an algorithm to select the most appropriate models from the diverse set of candidates that collectively address the uncertainty inherent in obesity dynamics.

The proposed methodology has several potential applications. In this paper, we apply it to the study of obesity in the region of Valencia, Spain, using available data to inform the model. Likewise, strategies and campaigns to reduce obesity can be introduced into the model to simulate their effect over time. However, the approach can be adapted and applied to other contexts where data are available, but knowledge about the phenomenon to be studied may be limited or unavailable. The models obtained can be projected to estimate the evolution of the system into the future.

The rest of the paper is structured as follows. Due to the difficulty of having data related to the evolution of obesity, in Sect. 2, we make a brief description of an already existing model (Sect. 2.1), and we use it to the generation of an uncertainty-aware synthetic dataset (Sect. 2.2). The full methodology is presented in Sect. 3, including the description of DSGE (Sect. 3.1) and the selection algorithm (Sect. 3.4). The experimental results, shown in Sect. 4, are discussed in Sect. 4.1. Finally, in Sect. 5, we summarise the conclusions of the paper.

2 Model and uncertainty-aware synthetic dataset generation

Unfortunately, getting data on the number of normal-weight, overweight, or obese people is complicated. In the Community of Valencia, Spain, there is a regular health survey from which this information can be obtained, although it is only every five years. This inconvenience is extensible to many areas of research.

Thus, to generate a dataset with a sufficient number of data describing the evolution over short periods of the number of people with normal weight, overweight, and obese, taking into account the uncertainty of the results obtained by conducting surveys, we will use the model proposed in Santonja et al. (2012). With it, we will generate an uncertainty-aware complete dataset.

2.1 Short description of the model describing the obesity dynamics

When classifying people according to weight, it is common to use the Body Mass Index (BMI) as metric, which is defined by $BMI = weight/height^2$. It is considered all people with a BMI lower than 25 are of normal weight (N), those in the 25 to 30 range are considered overweight (S), and finally, those with a value higher than 30 are categorized as obese (O). The model works with the population of the Community of Valencia, Spain, between 24 and 65 years of age at a specific time instant t, so we consider the different subpopulations as $N(t)$, $S(t)$ and $O(t)$, representing the percentage that they constitute of the total population. As it was mentioned, based on Hill and Peters (1998), Christakis and Fowler (2007), Cohen-Cole and Fletcher (2008), and Leahey et al. (2011), it is possible to transmit among people both poor eating habits and sedentary lifestyles leading to weight gain, which we summarize as *unhealthy habits*. Furthermore, assuming the classical hypothesis that populations are homogeneously mixed, the following rules to describe the transmission dynamics of unhealthy habits leading to weight gain are proposed:

- We distribute the individuals who enter the system at age 24 among the populations according to the distribution at age 23, which we denote by N_0, S_0, and O_0.

- An individual from subpopulation N transits to S due to the transmission of unhealthy habits from individuals belonging to S or O through social contact, leading to the individual gaining weight and moving to O.
- In the event that an individual of S persists in having unhealthy habits, over time, may become part of the population O.
- By changing habits to healthier ones, such as dieting and physical exercise, it is possible for an individual to reduce weight and move from O to S or from S to N.

Then, the obesity dynamics is described by the following system of difference equations[1] (t, time in weeks, $t = 0$ corresponds to the last week of year 2000),

$$N(t+1) - N(t) = \mu N_0 - \mu N(t) - \beta N(t)(S(t) + O(t)) + \rho S(t), \tag{1}$$

$$S(t+1) - S(t) = \mu S_0 - \mu S(t) + \beta N(t)(S(t) + O(t)) - (\gamma + \rho)S(t) + \epsilon O(t), \tag{2}$$

$$O(t+1) - O(t) = \mu O_0 - \mu O(t) + \gamma S(t) - \epsilon O(t). \tag{3}$$

The variables used in the model are:

- $1/\mu$ denotes the time from when an individual turns 24 and enters the system until turns 65 and leaves it, measured in weeks.
- β is the transmission rate of unhealthy lifestyles.
- $1/\gamma$ is the average time it takes for an overweight person with unhealthy habits to become obese.
- $1/\rho$, represents the average time required to reach a normal weight, having previously been overweight and having changed to healthier habits.
- $1/\epsilon$ represents the average time at which an individual from O transits to S, that is, the rate at which obese individuals become overweight.

The initial conditions, the results of model calibration, and the estimation of the model parameters variability are gathered in Table 1. More details about how the model parameters are calibrated and how the parameter distributions are assigned can be found in Santonja et al. (2012).

Note that the sum of the three subpopulations, by definition, satisfies the following condition of constant population:

$$N(t) + S(t) + O(t) = 1. \tag{4}$$

The graph flow representing this system of equations can be seen in Fig. 1.

2.2 Generation of the synthetic datasets using the obesity model

Here, we use the model in Sect. 2.1 to generate a set of uncertainty-aware synthetic datasets. Figure 2 summarizes the process of obtaining it. We start taking a high number of samples (500 in this case) for each of the model parameters following the uniform distributions in column 3 of Table 1, except for O_0 and $O(t = 0)$, which are calculated using (4). We

[1] The original model in Santonja et al. (2012) is a system of differential equations. However, we consider here a system of difference equations for better computational implementation and treatment. The model parameters are the same for both the differential equation model and its discretization.

Table 1 Calibrated model parameters and model parameter distributions. $t = 0$ corresponds to the last week of year 2000. N_0, S_0 and O_0 are the distribution of individuals at $t = 0$ at age 23 (just before enter into the system)

Parameter	Deterministic value	Interval	Distribution
μ	$\frac{1}{2184} = 0.0004578$	–	–
β	0.001121	[0.0008970, 0.001345]	Uniform
γ	0.0003226	[0.0002581, 0.0003871]	Uniform
ϵ	0.0000137143	[0.00000433, 0.00003248]	Uniform
ρ	0.00012	[0.00006, 0.0002256]	Uniform
N_0	0.704	[0.69, 0.71]	Uniform
S_0	0.25	[0.23, 0.26]	Uniform
O_0	0.046	[0.039, 0.052]	Uniform
$N(t = 0)$	0.522	[0.507, 0.536]	Uniform
$S(t = 0)$	0.362	[0.347, 0.376]	Uniform
$O(t = 0)$	0.116	[0.106, 0.125]	Uniform

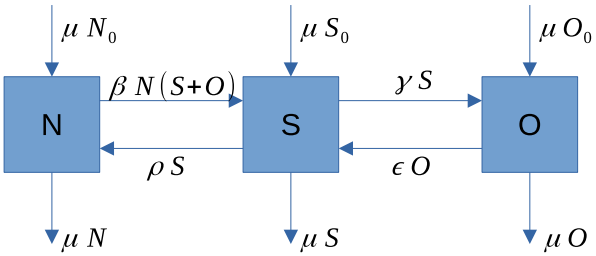


Fig. 1 Representation of the flow of individuals between subpopulations modeled in (1)–(3)

substitute each one of the 500 samples into the model (1)–(3) and run 500 simulations to obtain, for each one, the weekly evolution of the three subpopulations between the last week of 2000 and the last week of 2014, a total of 730 weeks.

The choice of 500 samples/model outputs is made with the idea that they are more than enough to capture all possible variations (uncertainty) that may occur in the model. However, in order to facilitate data handling and further work, it is desirable to reduce the number of samples/model outputs while still describing the uncertainty accurately.

Thus, to reduce the final number of samples, we take a set of $n, n = 10, 20, \ldots, 500$ model outputs, we calculate the mean and the 2.5 and 97.5 percentiles for each week, and we obtain the corresponding vectors with 730 elements, $m^n(N)$, $p^n_{2.5}(N)$, $P^n_{97.5}(N)$ for normal-weight N, $m^n(S)$, $p^n_{2.5}(S)$, $P^n_{97.5}(S)$ for overweight S and $m^n(O)$, $p^n_{2.5}(O)$, $P^n_{97.5}(O)$ for obese O subpopulations, respectively.

Now, we calculate the following composite error, $n = 10, 20, \ldots, 490$,

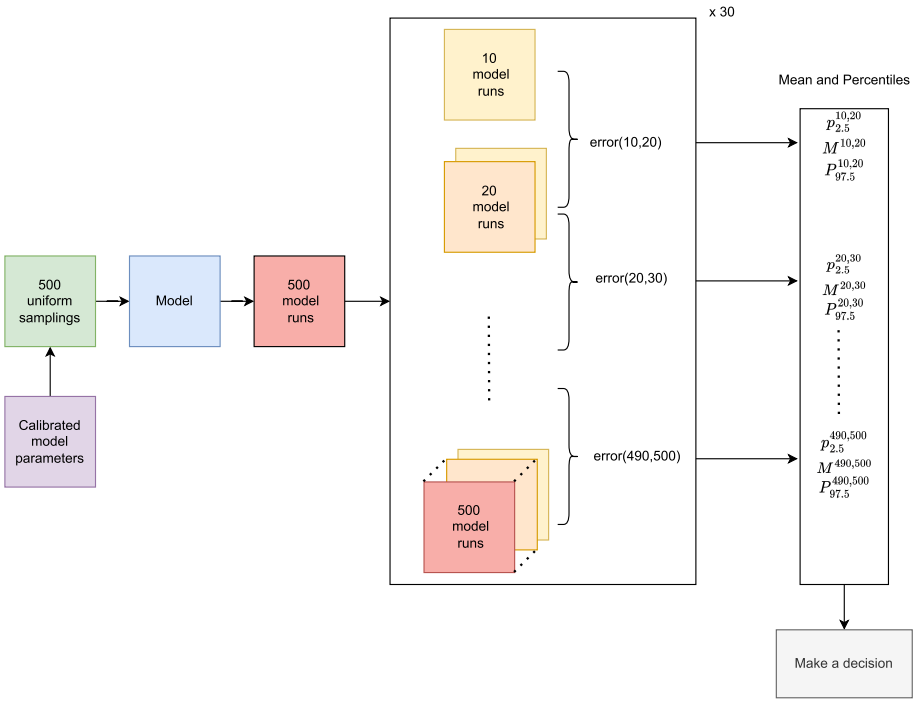


Fig. 2 Diagram of synthetic datasets generation. The stages represented are obtaining initial data by parameter sampling, using the three-equation obesity model to obtain the datasets, and selecting uncertainty-aware datasets

$$\begin{aligned} error(n, n+10) =& RMSE(m^n(N) - m^{n+10}(N) + RMSE(p_{2.5}^n(N) - p_{2.5}^{n+10}(N) \\ &+ RMSE(P_{97.5}^n(N) - P_{97.5}^{n+10}(N) \\ &+ RMSE(m^n(S) - m^{n+10}(S) + RMSE(p_{2.5}^n(S) - p_{2.5}^{n+10}(S) \\ &+ RMSE(P_{97.5}^n(S) - P_{97.5}^{n+10}(S) \\ &+ RMSE(m^n(O) - m^{n+10}(O) + RMSE(p_{2.5}^n(O) - p_{2.5}^{n+10}(O) \\ &+ RMSE(P_{97.5}^n(O) - P_{97.5}^{n+10}(O), \end{aligned}$$

where $RMSE$ denotes the root mean square error given by

$$RMSE((u_1, \ldots, u_p), (v_1, \ldots, v_p)) = \frac{1}{p}\sqrt{\sum_{i=1}^{p}(u_i - v_i)^2}.$$

The more model outputs are included, the more the composite error is reduced.

We repeat the above process 30 times, randomly reordering each time the model outputs. Then, we will have 30 $error(n, n+10)$ for each $n = 10, \ldots, 490$. For each group of 30 errors, we calculate the mean and 95% confidence interval. The idea of performing these 30 repetitions is to ensure that the selection we will make of the model output is independent of the randomness of the process and the order in which they have been generated.

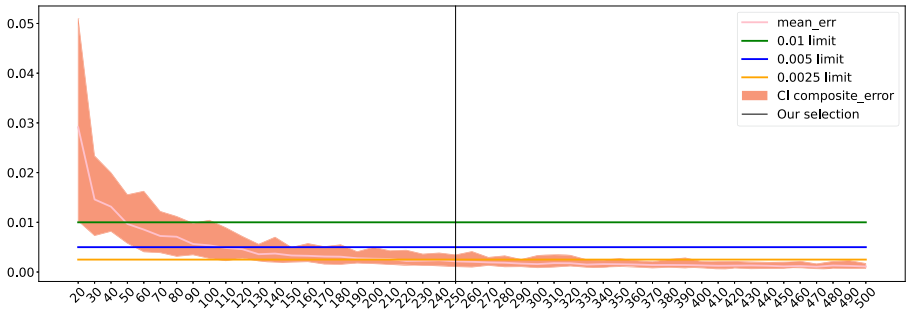


Fig. 3 Evolution of the mean and confidence interval of the composite error for the 30 runs. In order to find an equilibrium between future computational costs and low error, we take 250 simulations (black vertical line)

Figure 3 shows the evolution of the mean and confidence interval of the composite error for the 30 runs as a function of the number of model outputs used. In addition, three horizontal lines with values 0.01, 0.005, and 0.0025 have also been included to facilitate data visualization. To maintain a low error and also minimize the number of model outputs so as not to incur a high computational cost when performing the experiments, we have chosen to take a total of 250 model outputs to find an equilibrium between future computational costs and low error, represented in Fig. 3 by the vertical black line. Each of the 250 model outputs, each one with the 730 weekly values for normal weight, overweight, and obese populations, corresponds to a dataset we will work with next.

For later use, we establish the notation that we will use for these 250 datasets. Each dataset contains a time series of 730 elements for normal weight, overweight, and obese, and we denote them, for $i = 1, 2, \ldots, 250$, as

$$\begin{array}{ll} \text{normal weight} & N_1^i, N_2^i, \ldots, N_{730}^i, \\ \text{overweight} & S_1^i, S_2^i, \ldots, S_{730}^i, \\ \text{obese} & O_1^i, O_2^i, \ldots, O_{730}^i. \end{array} \tag{5}$$

This way, we have the 250 uncertainty-aware synthetic datasets we will work with. The successive differences of elements in the above time series in (5), is denoted, for $t = 1, 2, \ldots, 729$ and $i = 1, 2, \ldots, 250$, as

$$\Delta N_t^i = N_{t+1}^i - N_t^i, \qquad \Delta S_t^i = S_{t+1}^i - S_t^i, \qquad \Delta O_t^i = O_{t+1}^i - O_t^i. \tag{6}$$

In Fig. 4, we can see the 95% confidence band generated by the 250 time series in (5). The mean and the 2.5 and 97.5 percentiles that determine the 95% confidence interval (CI95%) are calculated as follows: for $t = 1, 2, \ldots, 730$,

- take the 250 values $N_t^1, \ldots, N_t^{250}$ and calculate their mean m_t^N, their percentile 2.5 p_t^N and their percentile 97.5 P_t^N,
- take the 250 values $S_t^1, \ldots, S_t^{250}$ and calculate their mean m_t^S, their percentile 2.5 p_t^S and their percentile 97.5 P_t^S,
- take the 250 values $O_t^1, \ldots, O_t^{250}$ and calculate their mean m_t^O, their percentile 2.5 p_t^O and their percentile 97.5 P_t^O.

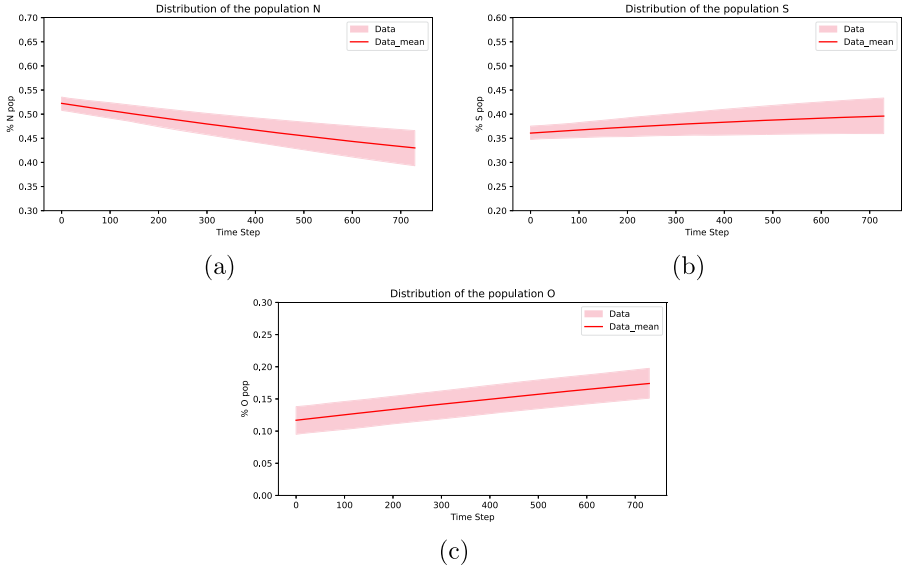


Fig. 4 The mean time series (red line) and the 95% confidence band (pink area) calculated in (7) from the 250 datasets time series in (5) for normal weight $N(t)$ (**a**), overweight, $S(t)$ (**b**), and obese $O(t)$ (**c**) subpopulations. These are the uncertainty-aware synthetic datasets we will work with (color figure online)

This way, the 250 uncertainty-aware synthetic datasets provide a mean and a 95% confidence interval time series, drawn in Fig. 4, and denoted for each week $t = 1, 2, \ldots, 730$, as

$$\begin{array}{ll} \text{mean} & m_t = (m_t^N, m_t^S, m_t^O)^T, \\ \text{percentile 2.5} & p_t = (p_t^N, p_t^S, p_t^O)^T, \\ \text{percentile 97.5} & P_t = (P_t^N, P_t^S, P_t^O)^T, \end{array} \tag{7}$$

where v^T denotes the transpose of vector v.

3 Methods

In the previous section, we obtained uncertainty-aware datasets of 250 model outputs from model (1)–(3). Now we forget all about this model and focus only on datasets to build a model.

This is the paper's main idea: to build a model based on known evidence but considering possible terms that may describe unknown features of the phenomenon, using only the datasets.

For this, we employ Dynamic Structured Grammar Evolution (DSGE) (Lourenço et al. 2018), designing a custom grammar that respects the known properties of the phenomenon under study and allows the inclusion of terms that are not contradictory to known properties. Although we have data for the three subpopulations, N, S, and O, we know that $O = 1 - N - S$. Also, since the individuals in N only enter because they enter the system, and if

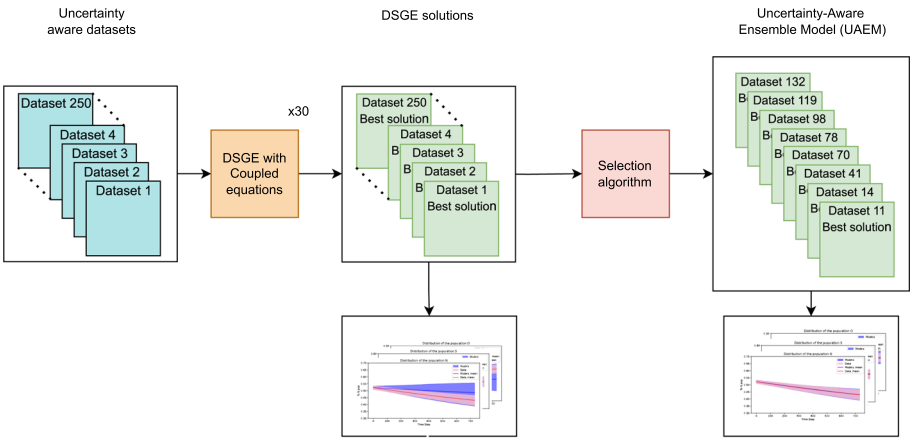


Fig. 5 In this figure, we show the workflow representation describing the obtention of the uncertainty-aware models. First, we apply DSGE 30 times to each dataset of overweight population S, taking the best-obtained model among the 30. As we can see in the bottom central image, the uncertainty described by these 250 models is far from matching the one shown in Fig. 4. To improve this matching, we use an uncertainty-aware selection algorithm to take a subset of models (ensemble model) that better describe the uncertainty drawn in Fig. 4 for the three populations, S, N, and O, using RMSE as the fitness function

they leave, or leave the system or move to S, once the expression of S is determined, that of N can be determined very simply.

Then, DSGE will be run 30 times to obtain 30 models for subpopulation S that accurately describe each one of the 250 datasets. We take the best of these 30 models. In the end, we will have 250 best models, one for each dataset. The selected fitness function measures the RMSE between the dataset values of subpopulation S and the predictions generated by the equations of DSGE. Finally, among these 250, we apply a *ad-hoc* uncertainty-aware selection algorithm to find out among the 250 models for subpopulation S which ones allow, not only to subpopulation S but also for populations N and O, to capture the data uncertainty drawn in Fig. 4 (ensemble model). Figure 5 outlines the methodology of the study.

3.1 Dynamic structured grammatical evolution

To generate the mathematical expressions for our obesity model, we employed a method called dynamic structured grammatical evolution (DSGE) (Lourenço et al. 2018). This approach is an improved version of the concept of Grammatical Evolution (GE).

DSGE is an evolutionary population-based algorithm, i.e. works with a set of representations of solutions (population) that iterates (evolves) over generations guided by a fitness function. The representations of the solutions are a list of integers that are decoded using a grammar to construct mathematical expressions. The algorithm has as inputs the dataset used for evaluation (6), the grammar, and the parameters of the algorithm. The output will be the mathematical expressions that best represent the data. The evolutionary process works as follows:

- Initialization (Step 0): A random population of lists of integer values is generated according to the parameters of the DSGE configuration.

- Decoding phase (Step 1): This step involves decoding the genetic representation, i.e., the list of integers, of each solution (individual) using the provided grammar and generating the necessary mathematical expressions
- Fitness computation (Step 2): when the solutions are decoded, we have expressions representing the successive differences in subpopulations N and O, that is, $\Delta N(t) = N(t+1) - N(t)$ and $\Delta O(t) = O(t+1) - O(t)$. Using them, we simulate the evolution of the groups in the population during the 730 weeks and compare the results with the input data. As the objective is to reproduce the dynamics of the problems as close as possible to the actual value, we measure the Root Mean Squared Error (RMSE) between the simulated data and the input data. The simulation is performed using *Iterative numerical evaluation* (INE), i.e., we iterative generate the 730 values of data using the one in the previous simulation step.
- Selection (Step 3): This phase determines which individuals from the population will form the parent set for producing the next generation. It uses a selection mechanism that favors individuals with lower error values, promoting the retention of more promising solutions.
- Crossover (Step 4): In the crossover phase, genetic material is exchanged between selected parent individuals, leading to the creation of new offspring. This process mimics the biological concept of genetic recombination and promotes the exploration of different solution combinations.
- Mutation (Step 5): The mutation phase introduces specific changes to individual solutions by altering their genetic representation. This process helps maintain genetic diversity within the population and enables the exploration of potentially novel solutions.

This iterative process is repeated until the specified number of generations is reached, allowing the population to evolve and improve solutions over time. The complete evolutionary process enables the discovery and refinement of expressions that capture the dynamics of obesity-related subpopulations.

3.2 Introducing knowledge through grammars

One of the advantages of using DSGE is the use of grammars to generate the solutions. In the context of Grammatical Evolution (GE), a grammar is a crucial component that defines the rules and structures for generating solutions to problems. With it, we can introduce restrictions or expert knowledge in the search. Figure 6 shows the grammar used in this paper, represented in Backus-Naur Form (BNF). Before explaining the particularities associated with the problem, let us introduce briefly the semantics of it. BNF is a notation technique used to express the grammar of a language in a formal way. It is commonly used in computer science for defining the syntax of programming languages, protocols, and data formats. Understanding the components and structure of a BNF grammar is essential for parsing and generating strings within a language. The building blocks of BNF grammars are:

- Symbols: There are two types, non-terminal and terminal symbols. Non-terminals are abstract symbols that can be expanded into sequences of other non-terminal or terminal symbols. They are represented by names enclosed in angle brackets <expression>. Terminals are the actual tokens of the language. They are the basic, indivisible elements that appear in the strings generated by the grammar. In other terms, the variables and operators that express the equations.
- Production Rules: They define how a non-terminal symbol can be replaced (or expanded) into a sequence of terminal and/or non-terminal symbols.

- Choices: A non-terminal symbol can often be expanded in more than one way. This is represented by listing multiple options, separated by a vertical bar (|), indicating a choice.
- The *<Start Symbol>* is a special non-terminal symbol designated as the initial symbol. This symbol represents the whole language or the main structure being defined. All strings in the language are derived by starting with this symbol and applying production rules.

As the grammar defines the structure of the equations the algorithm will obtain, we can also incorporate information into the grammar to favor some types of equations. In order to maintain the total proportion of the population, ΔS has been calculated as a function of $\Delta N(t)$ and $\Delta O(t)$. DSGE will use $\Delta S(t)$ as the objective to be adjusted; that is, we will try to calibrate the variation of the subpopulation S. This decision was made on the basis that this subpopulation is the center of the system under study. An individual moves to S, leaves S, or leaves the system. So the value of $\Delta S(t)$ will be the sum of the changes perceived in the other two subpopulations with opposite signs; this statement can be expressed as:

$$\Delta S = -(\Delta N + \Delta O) \tag{8}$$

This information is incorporated in the grammar of Figure 6 in the Start Symbol:

```
<func>::= ΔS(<ΔN>,<ΔO)>,k)
```

We will also search for expressions that take into account some terms of the equations obtained in Sect. 2, and Eqs. (1)–(3). If we translate the equations to the symbolic regression (SR) problem, we obtain the form of Eqs. (9)–(11):

$$\Delta N_{SGE}(t) = \mu N_0 + \Xi_N(t) - \mu N(t) \tag{9}$$

$$\Delta S_{SGE}(t) = \mu S_0 + \Xi_S(t) - \mu S(t) \tag{10}$$

$$\Delta O_{SGE}(t) = \mu O_0 + \Xi_O(t) - \mu O(t) \tag{11}$$

where $\Xi_N(t)$, $\Xi_S(t)$, and $\Xi_O(t)$ were the expressions obtained by DSGE (SR of the data). This is incorporated in the grammar in the non terminal lines:

```
<ΔN>::= <mu>* <N0> + <expr2N> + <expr2N>+ <expr2N> -<mu>* N
<ΔO>::= <mu>* <O0> + <expr2O> + <expr2O>+ <expr2O> -<mu>* O
<mu>::= 0.0004578
<N0>::= N0
<O0>::= O0
```

The intentional repetition of the terms $< expr2N >$ and $< expr2O >$ in the initial lines of the grammar reflects a deliberate design strategy informed by our experience with GE. Its inclusion empowers the algorithm to navigate the fitness landscape effectively, circumventing local minima and improving overall solution quality.

The next section of the grammar to highlight are the variables, differentiated according to the subpopulation and biased according to the knowledge about the system. Thus, the signs are prefixed depending on whether the variable or the product of the variable is considered incoming or outgoing from the subpopulation, as stated in the model in Sect. 2.1. In addition, we incorporate the following information through the grammar:

- The contagion of unhealthy lifestyles takes place through encounters of overweight (S) or obese (O) people with people of normal weight (N). Therefore, terms containing $N \times S$

and $N \times O$ can be generated by the grammar to permit the transition of individuals from N to S.

- Any individual can transit from S to O by increasing his or her unhealthy habits. And also the opposite if the individual turns to healthy habits. S can transit to N by changing the habits to healthier ones, such as going on a diet or doing exercise. Furthermore, terms describing the people entering or leaving the system to/from any subpopulations should also be contemplated.
- Terms with a higher degree, for instance, $S \times O^2$, are not allowed because there is no natural interpretation in the context of the problem. However, some freedom in the terms generation should be allowed because they may not be contradictory with the behavior of the problem.
- Also, the people in N who leave N, move to S to keep the total population constant. Thus, the terms that add in S, except those corresponding to individuals who enter or leave the system, are those that subtract in N. In this way, the equation generated by the algorithm for the subpopulation S allows us to obtain the corresponding equation in N, and from both, the one for O is obtained as $1 - N - S$. Then with the equation of S, the remainder are determined.

All this information is incorporated in the grammar in the lines

```
<expr2N>::= (<exprN> + <exprN>) | (<cte> * <varN>)
<exprN>::= <varN> | (<cte> * <varN>)|<expr2N>
<expr2O>::= (<exprO> + <exprO>) | (<cte> * <varO>)
<exprO>::= <varO> | (<cte> * <varO>)|<expr2O>
<varN>::= -(N)|-(N*S)|-(N*O)|(S)|(S*N)
<varO>::= S|S*O|-(O)|-(S*O)|-(N*O)
```

The rest of the grammar is devoted to the generation of constants.

3.3 Evaluation of the tentative models

Figure 6 presents an example of a representation of our context-free grammar used for defining the syntax of coupled equations in the context of obesity modeling based on the genotype of each individual. The genotype comprises a sequence of numbers (*alleles* in this context), with each number serving as a basis for generating a sequence of terminals and non-terminals derived from the grammar file. The process starts with the non-terminal symbol $< func >$, which is replaced by the expression $\Delta S(< \Delta N >, < \Delta O >), k)$. In Fig. 7, we can see the steps to develop ΔN as an example. The green boxes denote the grammar rules. The blue squares show the intermediate expressions. The process commences by applying the initial grammar rule alongside the first number within our genotype. Subsequently, the progression involves decoding the generated non-terminals sequentially using the corresponding genotype numbers (orange ellipses) until a collection of terminals forms the phenotype (the final expression).

Applying analogous procedures, the algorithm proceeds to compute ΔO, culminating in the assembly of ΔS in a coupled format, illustrated in Fig. 8. Subsequently, the model's outcomes are compared against the dataset values, and the model's error is quantified, defining the fitness of the individual. This fitness metric is the criterion enabling the evolutionary algorithm to discern and select the most optimal models (individuals) for generating subsequent generations.

```
# Model expression
<func>::= ΔS(<ΔN>,<ΔO)>,k)

<ΔN>::= <mu>* <N0> + <expr2N> + <expr2N>+ <expr2N> -<mu>* N
<ΔO>::= <mu>* <O0> + <expr2O> + <expr2O>+ <expr2O> -<mu>* O
<mu>::= 0.0004578
<N0>::= N0
<O0>::= O0

<expr2N>::= (<exprN> + <exprN>) | (<cte> * <varN>)
<exprN>::= <varN> | (<cte> * <varN>)|<expr2N>
<expr2O>::= (<exprO> + <exprO>) | (<cte> * <varO>)
<exprO>::= <varO> | (<cte> * <varO>)|<expr2O>
<varN>::= -(N)|-(N*S)|-(N*O)|(S)|(S*N)
<varO>::= S|S*O|-(O)|-(S*O)|-(N*O)

<cte>::= <base>*Math.pow(10,<sign><exponent>)
<base>::= 1|2|3|4| ... |99
<exponent>::= 1|2|3|4|5|6|8|9
<sign>::= +|-
```

Fig. 6 Grammar applied in our work for decoding the individuals

3.4 Construction of the uncertainty-aware ensemble model (UAEM) using a selection algorithm

The objective now is to propose an algorithm to select, among the obtained models, a subset that will make up the uncertainty-aware ensemble model that will capture as accurately as possible the data uncertainty, that is, when the mean and the 95% confidence interval of the models will be as close as possible to the mean and the 95% confidence interval of the data.

At this point, we have 250 models $M_i(t)$, $i = 1, \ldots, 250$ obtained using DSGE, where

$$M_i(t) = (N_i(t), S_i(t), O_i(t))^T,$$

is a vector with three components, the percentage of normal weight, overweight, and obese. With these models, we can evaluate their corresponding model outputs $M_i(1), \ldots, M_i(730)$, at the weeks $t = 1, 2, \ldots, 730$, that is:

$$\begin{array}{lll} \text{Index} & \text{Models} & \text{Output} \\ 1 & M_1(t) & \Theta(1) = (M_1(1), \ldots, M_1(730)), \\ 2 & M_2(t) & \Theta(2) = (M_2(1), \ldots, M_2(730)), \\ \vdots & \vdots & \vdots \\ 250 & M_{250}(t) & \Theta(250) = (M_{250}(1), \ldots, M_{250}(730)). \end{array} \tag{12}$$

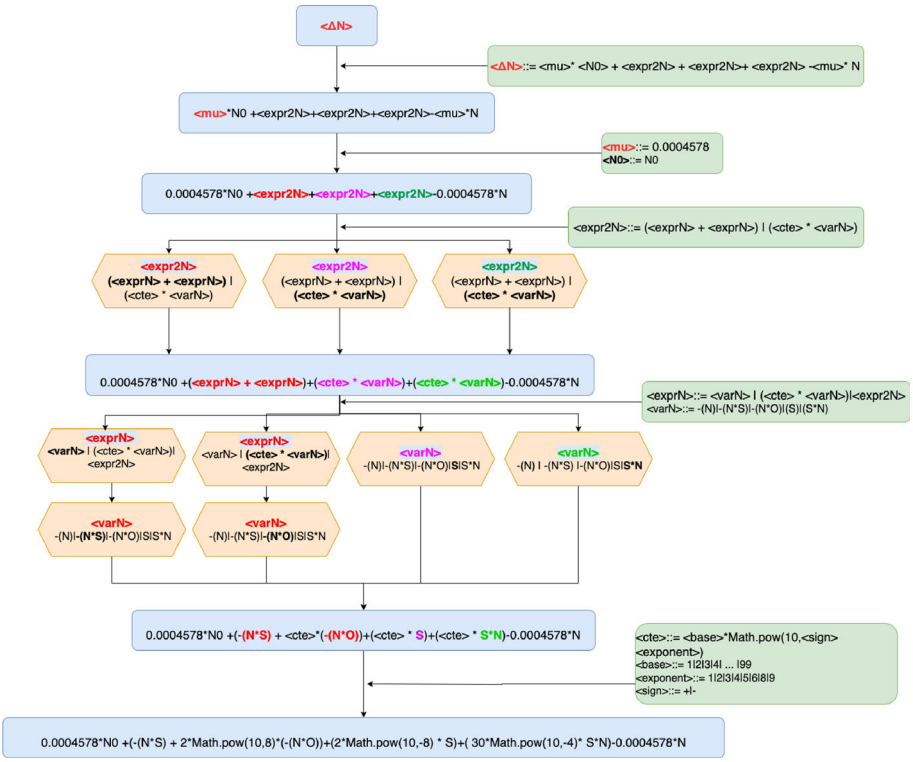


Fig. 7 Flowchart for a hypothetical example of the decoding process of ΔN

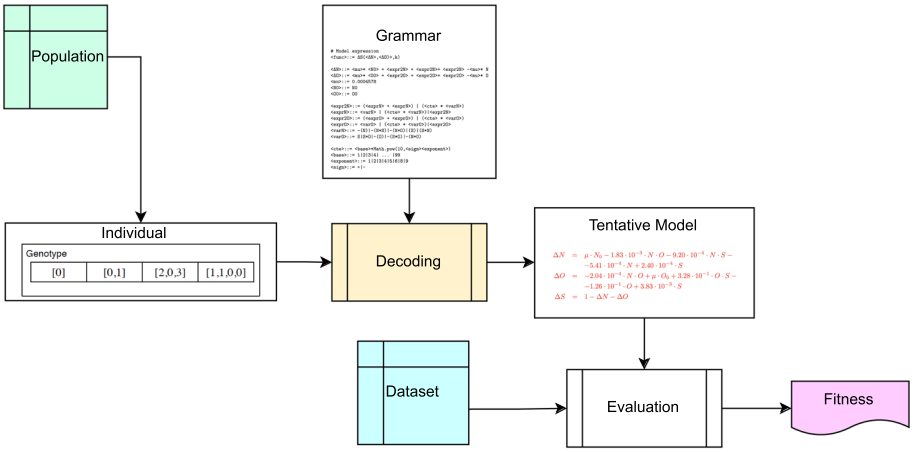


Fig. 8 Flow diagram of the decoding and evaluation process

If we consider $I_h \subseteq I = \{1, \ldots, 250\}$ a subset of indexes of I, taking the rows $\Theta(i)$, $i \in I_h$, for each of the three elements in each column we can calculate the corresponding percentiles 2.5, 97.5 and the mean, and denote them, for $t = 1, 2, \ldots, 730$, as

$$\begin{array}{ll} \text{mean} & m(t)^{I_h} = (m_N^{I_h}(t), m_S^{I_h}(t), m_O^{I_h}(t))^T, \\ \text{percentile 2.5} & p(t)^{I_h} = (p_N^{I_h}(t), p_S^{I_h}(t), p_O^{I_h}(t))^T, \\ \text{percentile 97.5} & P(t)^{I_h} = (P_N^{I_h}(t), P_S^{I_h}(t), P_O^{I_h}(t))^T. \end{array} \tag{13}$$

Now, taking into account the mean and confidence intervals of the datasets defined in (7), we define here explicitly the error function as:

$$\begin{array}{rl} F(I_h) = & \sum_{t=1}^{730} \text{RMSE}(m(t)^{I_h} - m_t) \\ & + \text{RMSE}(p(t)^{I_h} - p_t) \\ & + \text{RMSE}(P(t)^{I_h} - P_t). \end{array} \tag{14}$$

Thus, the goal consists of finding a subset $I_{h^*} \subseteq I$ (ensemble model) such that the $F(I_{h^*})$ will be as small as possible (uncertainty aware).

To achieve that, we use the UAEM selection algorithm based on the PSO (Marini and Walczak 2015), which is explained as follows.

1. Parameters of the algorithm.
   - n, the number of particles.
   - $ITMAX$, the maximum number of iterations.
2. Initialization.
   - Let $I_1, \ldots, I_n \subset \{1, \ldots, 250\}$ be the sets of indexes (particles), where $|I_1|, \ldots, |I_n| < 250$. Note that the number of elements of each I_i, $i = 1, \ldots, n$ may be different.
   - Calculate the particles fitnesses $F(I_i)$, $i = 1, \ldots, n$, F defined in (14).
   - We define the local best of each particle as $I_i^{localbest} = I_i$, $i = 1, \ldots, n$.
   - We define the global best $I^{globalbest}$ as the best (lowest value of F) of all the local best.
3. STEP 1 (particle update). For $i = 1$ to n, we have two possible options to update the particle:
   - With 10% probability, we generate randomly a new particle I_i with a random size in $\{1, \ldots, 250\}$.
   - Otherwise, we define the auxiliary set S as the union of the elements of the I_i, the elements of $I_i^{localbest}$, the elements of $I^{globalbest}$ and a random amount of random numbers between 1 and 250. Then, we remove the repeated elements. The updated I_i will be made up of a random number of elements of S chosen at random.
4. STEP 2 (particle fitness calculation). For $i = 1$ to n, calculate the error $F(I_i)$ of the updated I_i.
5. STEP 3 (updating the local best and the global best) For $i = 1$ to n.
   - The local best of $I_i^{localbest}$ is updated if the new I_i has a fitness $F(I_i)$ less than $F(I_i^{localbest})$.
   - In the same way, if the local best has been updated, we check if $I^{globalbest}$ can be updated if I_i has a fitness $F(I_i)$ less than $F(I_i^{globalbest})$.

6. STEP 4 (end criterion) The process finishes when ITMAX iterations have been reached, and the algorithm returns $I^{globalbest}$. Otherwise, go to STEP 1.

The $I^{globalbest}$ returned by the algorithm will be a subset of $\{1, \ldots, 250\}$ whose fitness error, $F(I^{globalbest})$ will be the smallest found.

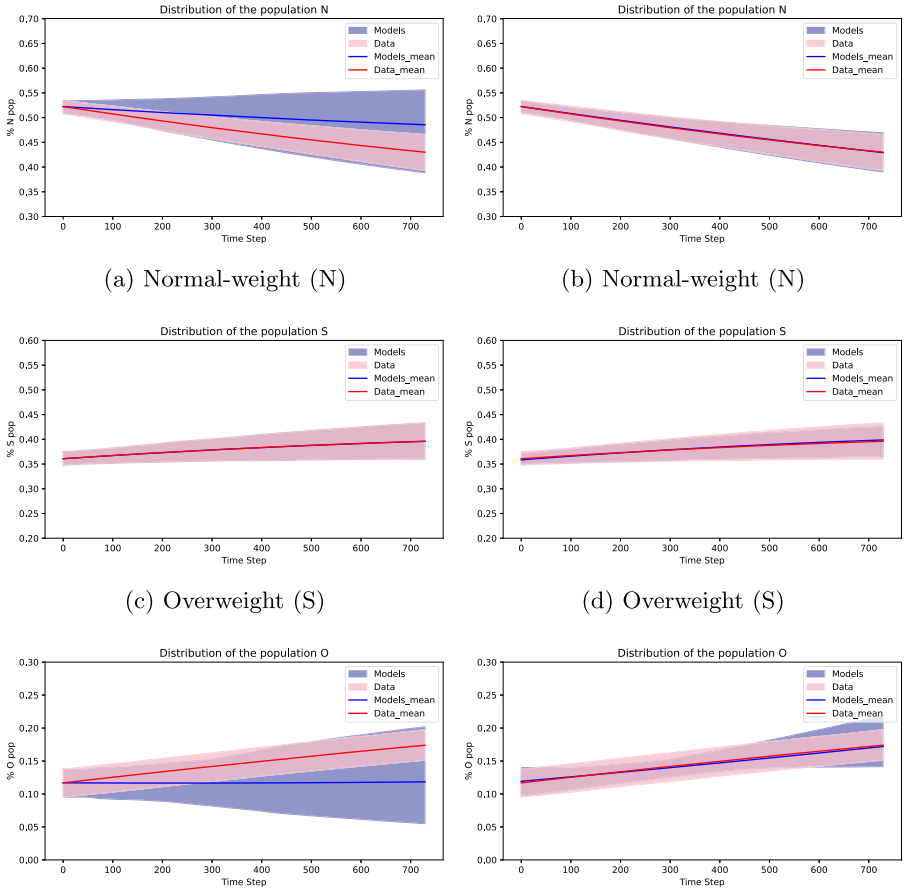
4 Experimental results

This study encompasses two distinct phases of results. Initially, the application of DSGE focused primarily on subpopulation S to evaluate the performance of candidate solutions. Subsequently, a second phase involved the utilization of the UAEM selection algorithm, which considered the performance of the selected solution set in predicting the evolution of all three subpopulations. The left column of Fig. 9 illustrates the outcomes: the pink region represents the corresponding interval obtained from the real data that is also drawn in Fig. 4, while the blue region represents the confidence interval between the upper and lower bounds of the data generated using the 250 models obtained using DSGE. Although the data obtained from the 250 models cover the uncertainty-aware datasets, they do not adequately fit the cases of N and O. We can see a progressive discrepancy between the two averages that increases as the prediction time horizon increases.

This discrepancy arises because the model was based only on values from the subpopulation S for fitting, and more guidance is necessary for the other two objectives. Since the obtained solutions do cover the desired space, the subsequent step involves the selection of expressions that faithfully represent the data. To achieve this, we employed the UAEM selection algorithm to minimize the difference between the mean and confidence intervals of the simulations generated by a subset of the 250 models and the actual data. We run the UAEM selection algorithm 30 times for $n = 40, 60, 80, 120, 140$ particles and ITMAX$=$ 20,000, 30,000 iterations. This process yields a set of expressions that collectively fit the original dataset. We run the algorithm with all this variety of options with the aim at covering many possibilities of execution of the algorithm to finally select the best solution. We do it this way because the computational cost of the algorithm is not high, and does not justify a study of the performance of the algorithm, which is outside the scope of this study. Among all of these experiments, the best result $I_{h^*} = \{11, 14, 41, 70, 78, 98, 119, 132\}$ was obtained with $n = 140$ particles, ITMAX$=$ 30,000 iterations, and a fitness error 0.0293171. These 8 models constitute the uncertainty-aware ensemble model, as can be seen in the right column of Fig. 9. This column is similar to the left column, but instead of employing the 250 obtained models, we utilize the 8 returned by the UAEM selection algorithm. As depicted, both the mean and the 95% confidence intervals of the data and the ensemble model align much better for subpopulations N and O than in the left column. The results obtained from the coupled equations closely approximate the values of the dataset, indicating the efficacy of our approach in addressing such problems.

4.1 Discussion

After applying UAEM selection algorithm (Sect. 3.4) to curate a subset of solutions capable of efficiently covering the uncertainties inherent in the original 250 uncertainty-aware synthetic datasets, we obtained 8 solutions listed in Table 2. This table shows each solution identified by an ID and differentiates between the expression for $\Delta N = N(t+1) - N(t)$ and $\Delta O = O(t+1) - O(t)$. We can use those expressions to extract $\Delta S = S(t+1) - S(t) = -(\Delta N + \Delta O)$,



(a) Normal-weight (N) (b) Normal-weight (N)

(c) Overweight (S) (d) Overweight (S)

(e) Obese (O) (f) Obese (O)

Fig. 9 Comparison of the mean and confidence intervals of the 250 datasets, as shown in Fig. 4 (pink), with: (left column) the mean and confidence intervals of the 250 models obtained from the DSGE (blue); (right column) the ensemble model (8 out of the 250 original models) obtained after applying the UAEM selection algorithm (color figure online)

as indicated in Eq. (8). According to Sect. 3.2, certain elements have been prefixed using grammatical rules, like the terms accompanied by $\mu = 0.0004578$.

Now, we analyze the results in Table 2, comparing them with the original model (1)–(3) proposed in Santonja et al. (2012) and recalling the hypothesis used to build this model. Specifically, we are going to look at the terms that the grammar has constructed but do not appear in the original model (1)–(3). Our task here will be to justify the appearance of these terms, either because they do not contradict the hypotheses or complement them.

First, we look at the expressions of ΔN:

1. The first equation (1) in the original model contains a non linear term involving SN and SO. Only solution IDs 1 and 2 contain both terms. The remainder contain only one of both. The differences among the expressions of the solutions may show the variability in

Table 2 Expressions obtained after the UAEM selection algorithm. ΔS can be obtained as $-(\Delta N + \Delta O)$, see Eq. (8)

ID	Expresiones
1	$\Delta N = \mu N_0 - \mu N - 0.00000179NO - 0.00091NS$
	$\Delta O = \mu O_0 - \mu O + 0.00186OS - 0.00029099O$
2	$\Delta N = \mu N_0 - \mu N - 0.004000015NO + 0.000199NS$
	$\Delta O = \mu O_0 - \mu O + 2.62OS - 1.00027O + 0.029S$
3	$\Delta N = \mu N_0 - \mu N - 0.002706NS + 0.00088S$
	$\Delta O = \mu O_0 - \mu O + 0.000782909OS + 0.0000000004S$
4	$\Delta N = \mu N_0 - \mu N - 0.00163NS + 0.0000029S$
	$\Delta O = \mu O_0 - \mu O - 0.00032NO - 0.00051097OS - 0.00000295O + 0.00049049S$
5	$\Delta N = \mu N_0 - \mu N - 0.0034NO - 0.00048N + 0.00042S$
	$\Delta O = \mu O_0 - \mu O - 0.00037NO - 0.00002968OS - 0.00019978O + 0.00043214S$
6	$\Delta N = \mu N_0 - \mu N - 0.00131NS - 0.00000399N$
	$\Delta O = \mu O_0 - \mu O - 0.00000045NO + 0.000052OS + 0.00025071S$
7	$\Delta N = \mu N_0 - \mu N - 0.0072NO - 0.00018N + 0.00062S$
	$\Delta O = \mu O_0 - \mu O - 0.000163NO + 0.002799OS$
8	$\Delta N = \mu N_0 - \mu N - 0.001NS + 0.00000098S$
	$\Delta O = \mu O_0 - \mu O - 0.00078NO + 0.000430973S$

the influence of the populations of S or O on the decision of the people in N to change their habits.

2. Solution IDs 5, 6, and 7 have a negative term in N. This term is missing in the Eq. (1). It may indicate that there are individuals who autonomously acquire unhealthy habits and, as a consequence, gain weight and become S. Although the general hypothesis is that the transition from N to S is due to transmission of unhealthy habits, it seems complementary the possibility that some individuals do it voluntarily, even more so if we realize that the coefficients of these terms are of a significantly smaller magnitude than the corresponding to the nonlinear terms.

Now, looking at the expressions of ΔO, all the solutions contain a non linear term when non linear terms are missing in Eq. (3) of the original model.

1. Solution IDs 4, 5, 6, 7, and 8 have a negative term in NO. This term may show that people in N are able to transmit healthier habits leading people to move from O to S. Although the original model assumes that weight loss due to changing habits for healthier ones is a personal decision (linear term), the constructed model indicates that it may also occur in some cases due to the transmission of habits.
2. Now, we deal with the term OS appearing in all the solutions except the ID 8.
   - In solution IDs 1, 2, 3, 6, and 7, this term is positive. It shows the possibility that people in S move to O because people in O transmit more unhealthy habits leading people in S to gain weight.
   - However, in solutions 4 and 5, this term is negative, showing that people in S may transmit to people in O healthier habits than the usual for people in O and move to S.

As mentioned before, the original model assumes linear terms for these transitions, however, the model our algorithms built shows that these transitions may also occur in some cases due to the transmission of habits.

Summarizing, the solutions in Table 2 are in agreement with the known hypothesis and do not contradict them. Also, the new terms suggest that it may be interesting to study possible causes of transitions between populations, extending the transmission of habits also to lose weight, as well as the autonomous decision to gain/lose weight.

5 Conclusion

In this work, we develop a process such that, from datasets of the evolution of a phenomenon with uncertainty and having limited knowledge of this phenomenon, we build a set of dynamic models based on a system of difference equations describing the datasets and their uncertainty. Furthermore, these models are explainable and non-contradictory to the existing knowledge of the problem.

The primary technique employed is DSGE. When DSGE calculates the models, we apply a selection algorithm to reduce the number of models while improving the fit to the datasets.

Here are the key findings from our study:

- DSGE proved to be a valuable tool as it allowed us to incorporate the knowledge about the system into our models. This meant we could adjust our models based on prior evidence, making them more accurate.
- The use of sparse identification simplifies the problem. This technique focuses on the idea that in real-world systems, only a few factors play a crucial role in explaining how things work.
- Our models are not just accurate; they are also explainable. This means that they are easy for humans, especially experts in the area, to understand. The terms in the models respond to identifiable processes that can be described in the context of the studied phenomenon. Rather than having a *black box* model where you can only see the output, our models allow experts to identify the elements governing the system.
- The terms in the model not supported by evidence but non-contradictory with them may suggest new lines of research, as it happens in other areas.

In this study, we applied our methodology to a synthetic dataset that we created to account for a wide range of uncertainties. However, the real promise of our approach lies in its ability to derive equations that describe the dynamics of a real-world system. We envision a future where we can create interpretable models from a limited dataset and minimal prior knowledge of the underlying laws of the system.

It should be noted that we have been able to simplify a system of three equations into a single equation using the population conservation principle and individuals' structured movement between subpopulations. In more complex models where this kind of simplification cannot be done, it would be interesting to face this problem with a multiobjective approach, developing a model that simultaneously adjusts the coupled equations to capture the evolution of all the populations. Such a proposal would definitely increase the computational complexity and require optimization techniques oriented explicitly to multiobjective problems.

In essence, our work aims to make complex systems more transparent and understandable, offering a valuable tool for experts and decision-makers in various fields, particularly in the medical domain.

Acknowledgements This work has been supported by the grants PID2020-115270GB-I00, PDC2022-133429-I00 and PID2021-125549OB-I00 funded by MCIN / AEI/ 10.13039 / 501100011033 and by European Union Next GenerationEU / PRTR.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability All the necessary data are included in the paper.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

Centers for Disease Control and Prevention (CDC) (2022) CDC's Division of Nutrition, Physical Activity, and Obesity (DNPAO), "About overweight & obesity," https://www.cdc.gov/obesity/about-obesity/index.html. Accessed 12 May 2022

Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. New Engl J Med 357(4):370–379

Cohen-Cole E, Fletcher JM (2008) Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. J Health Econ 27(5):1382–1387

Evangelista AM, Ortiz AR, Ríos-Soto KR, Urdapilleta A (2004) USA the fast food nation: Obesity as an epidemic. T-7, MS B284, Theoretical Division, Los Alamos National Laboratory, NM 87545. https://mcmsc.asu.edusites/default/files/2024-08/MTBI%202004%20Obesity%20Epidemic%20Report.pdf

Frerichs LM, Araz OM, Huang TT-K (2013) Modeling social transmission dynamics of unhealthy behaviors for evaluating prevention and treatment interventions on childhood obesity. PLoS ONE 8(12):e82887. https://doi.org/10.1371/journal.pone.0082887

Hill JO, Peters JC (1998) Environmental contributions to the obesity epidemic. Science 280(5368):1371–1374

Hill AL, Rand DG, Nowak MA, Christakis NA (2010) Infectious disease modeling of social contagion in networks. PLoS Comput Biol 6(11):1–15. https://doi.org/10.1371/journal.pcbi.1000968

Leahey TM, LaRose JG, Fava JL, Wing RR (2011) Social influences are associated with bmi and weight loss intentions in young adults. Obesity 19(6):1157–1162

Lourenço N, Ferrer J, Pereira FB, Costa E (2017) A comparative study of different grammar-based genetic programming approaches. In: McDermott J, Castelli M, Sekanina L, Haasdijk E, García-Sánchez P (eds) Genetic programming. Springer International Publishing, Cham, pp 311–325

Lourenço N, Assunção F, Pereira FB, Costa E, Machado P (2018) Structured grammatical evolution: a dynamic approach. In: Ryan C, O'Neill M, Collins J (eds) Handbook of grammatical evolution. Springer International Publishing, Cham, pp 137–161

Lozano-Ochoa E, Camacho JF, Vargas-De-León C (2017) Qualitative stability analysis of an obesity epidemic model with social contagion. Discrete Dyn Nat Soc 2017:1–12. https://doi.org/10.1155/2017/1084769

Marini F, Walczak B (2015) Particle swarm optimization (PSO). A tutorial. Chemom Intell Lab Syst 149:153–165

Santonja F, Shaikhet L (2014) Probabilistic stability analysis of social obesity epidemic by a delayed stochastic model. Nonlinear Anal Real World Appl 17:114–125

Santonja F-J, Villanueva R-J, Jódar L, Gonzalez-Parra G (2010) Mathematical modelling of social obesity epidemic in the region of Valencia, Spain. Math Comput Modell Dyn Syst 16(1):23–34. https://doi.org/10.1080/13873951003590149

Santonja F-J, Morales A, Villanueva R-J, Cortés J-C (2012) Analysing the effect of public health campaigns on reducing excess weight: A modelling approach for the Spanish Autonomous Region of the Community of Valencia. Eval Progr Plan 35(1):34–39. https://doi.org/10.1016/j.evalprogplan.2011.06.004

Smith RC (2013) Uncertainty quantification: theory, implementation, and applications. SIAM, Cambridge

Wadhera D, Phillips EDC, Castillo-Chavez C, Safan M, Murillo AL (2016) Modeling eating behaviors: the role of environment and positive food association learning via a Ratatouille effect. Math Biosci Eng 13(4):841–855. https://doi.org/10.3934/mbe.2016020

WHO Discussion Paper: Draft recommendations for the prevention and management of obesity over the life course, including potential targets (2021) https://www.who.int/publications/m/item/who-discussion-paper-draft-recommendations-for-the-prevention-and-management-of-obesity-over-the-life-course-including-potential-targets. Accessed 12 May 2022