Frank Peter Pijpers
Korteweg-de Vries Institute
University of Amsterdam
f.p.pijpers@uva.nl

Inaugural Lecture

Empirical explorations of

On Wednesday 20 October 2021, Frank Pijpers delivered his inaugural lecture upon acceptance of the position of professor by special appointment in the field of complexity for official statistics at the Faculty of Science of the University of Amsterdam, on behalf of Statistics Netherlands (CBS).

Official statistics as a mirror of society

It is an important objective of the Complexity for Official Statistics chair to build a bridge between Statistics Netherlands (CBS), my main employer, and the knowledge and techniques that are developed within the academic discipline of complex systems. I therefore think it makes sense to tell something about each of those two separately; official statistics and complex systems, and only then describe my vision of what the two have to do with each other.

Before I get down to the task of doing all of that in a single public lesson, I think it's also good to share something about myself that is relevant. I believe that a good teacher starts by trying to understand what her or his audience already knows, or thinks they know, about the subject. This is certainly the case when such a teacher tries to use new knowledge to indicate the boundaries of something that that audience considers a truth. The reverse also applies: it is good if an audience knows something about the background, knowledge and experience of the researcher who teaches, and the way in which the researcher examines and views reality.

It is therefore important that you know that I was trained as an astrophysicist and that for years I have conducted research in theoretical or mathematical sub-areas within it. This means, among other things, that I tend to think that *how long* a phenomenon or a process lasts is only meaningful in relation to another time scale belonging to a related or relevant other process. Now you are forewarned that I may take giant strides through the blink of an eye that is human history, without always making a distinction between a decade and a millennium. In addition, this background also means that my world view revolves around physical processes and interactions rather than things. You may think that this mindset is not the best preparation for taking up employment at an agency whose task it is to count people and companies, and categorize and describe their characteristics, and neatly tabulate and publish the results. I hope I can convince you that this is an incomplete view of the purpose of a national statistical institute.

I will test your patience only briefly with history lessons on the origins of official statistics. The idea of a regular census certainly existed in the Roman Empire, and over the same era in China and India. The purpose of such censuses can be classified as “supporting public administration by providing facts about the population” as it is now. It is fair to say that public administration at that time may have limited itself mainly to levying taxes and raising armies: to put it bluntly it was about money and power. In addition to self-enrichment or glorification, however, certainly public works were built with

Frank Pijpers

a complex society

that tax revenue, such as roads, viaducts and aqueducts.

In this country, Statistics Netherlands has now been in existence for 122 years and over the years it has started to keep track of — and monitor — more and more things that are important for public policy. Not only are data collected and published about the population, but also about health and security, about education and leisure activities, about the environment and all kinds of economic activity, both nationally and internationally. As a result, each government ministry has its own specialized factual material at its disposal to use when making policy decisions. Incidentally, I prefer to use the term measurement data rather than the term factual material, but the latter term is simply in vogue within Statistics Netherlands. It is much more recent in origin that there is a further separation between the collection and publication of measurement data, by Statistics Netherlands, and the use of models, calibrated on that data, for calculating the (potential) consequences of public policy choices, through the Netherlands Bureau for Economic Policy Analysis (CPB) and the even more recently established Netherlands Institute for Social Research (SCP) and the Netherlands Environmental Assessment Agency (PBL).

It is debatable whether this division between Statistics Netherlands and CPB was partly motivated to demarcate territory between the then CBS director-general Idenburg on the one hand, and Jan Tinbergen on the other, who was working at Statistics Netherlands at the time [5]. Tinbergen subsequently became the first director of the CPB, and was later also the winner of the Swedish Riksbank Prize in Economic Sciences; a prize that is equated with the Nobel Prizes. However, there is also a substantive reason for making this separation, just as a distinction is made between theoretical and experimental physics. In the case of the CBS-CPB distinction, this division explicitly prevents the impression that the measurement data might have been given a political slant. The consequences of the political choices presented in the national budget on Budget Day, are projected by calculations of the CPB through models. Statistics Netherlands' material thus explicitly retains a higher degree of objectivity and independence than might otherwise be the case. The intention is that Statistics Netherlands can hold up a mirror to the Netherlands as faithfully as possible.

Such a separation also has drawbacks, but to explain them I would first like to take you into the world of the science of complex systems.

The whole and the sum of its parts

Language use in a technical context, sometimes condescendingly referred to as jargon, makes use of very precise definitions if all goes well. That may seem fussy, but it is really necessary to enable, for example, that re-doing experiments or measurements is indeed carried out precisely and correctly. That reproducibility is fundamental and distinctive to real science. Technical language use therefore sometimes leads to misunderstandings if a certain word also has a daily, and much less precisely defined, use. The word 'complex' in complex systems is unfortunately such a word. In everyday usage it is often used as a synonym for complicated, or even synonymous with incomprehensible. Then it can happen that poorly documented and unstructured computer code is described as 'complex', when it is merely the result of bad habits and laziness. Every large organization has to deal with these kinds of problems, and Statistics Netherlands is certainly no exception, but that has nothing to do with the complex systems I want to talk about.

To paint a picture, I would like to give some analogies from nature that are expressions of complex systems, before taking a deeper dive into a more abstract, mathematical description. The first analogy is that of a termite mound. An individual termite is by itself an independent living organism. Considered as a unit among many members of the species, a single termite is not very intelligent: it reacts to external stimuli in quite restricted ways. Yet a large collection of termites is able to build structures on a scale that far exceeds their own size, and these are structures that also perform functions that allow a degree of control over the environment of that termite society, such as temperature control, which the single termite cannot possibly understand. Clearly, that termite society, the termite mound, is capable of behaviour that does not follow trivially from the behaviour of individual termites. The whole is more, can achieve more, than the sum of its parts. The construction of termite mounds, or the patterns that form in murmurations of starlings, are an expression of complex systems: this is summarily referred to by the term *emergent behaviour*.

We can also descend in scale level to that of atoms and molecules. Even with just the elements H, C, N, O and P, it is possible to build a huge variety of molecules. If those molecules are strung together in long chains, they can start to behave like self-reproducing machines. At this scale level it is indistinguishable whether we are looking inside humans or termites, or petunias, or whales [1]. There is no external controlling principle or overarching organisation that imposes an order of elements that is then reproduced. In fact, a long chain with the same sequence of elements can fold in different ways and thus acquire different chemical properties [10]. The possibility in itself that elements can form bonds naturally exists in the electron structure of individual atoms themselves. However, there are boundless possibilities to build structures out of this that don't even have to be chains. So whether, and which, chains are formed, and how they fold, is by no means fixed in the atoms: it is *self-organizing behaviour* of the system.

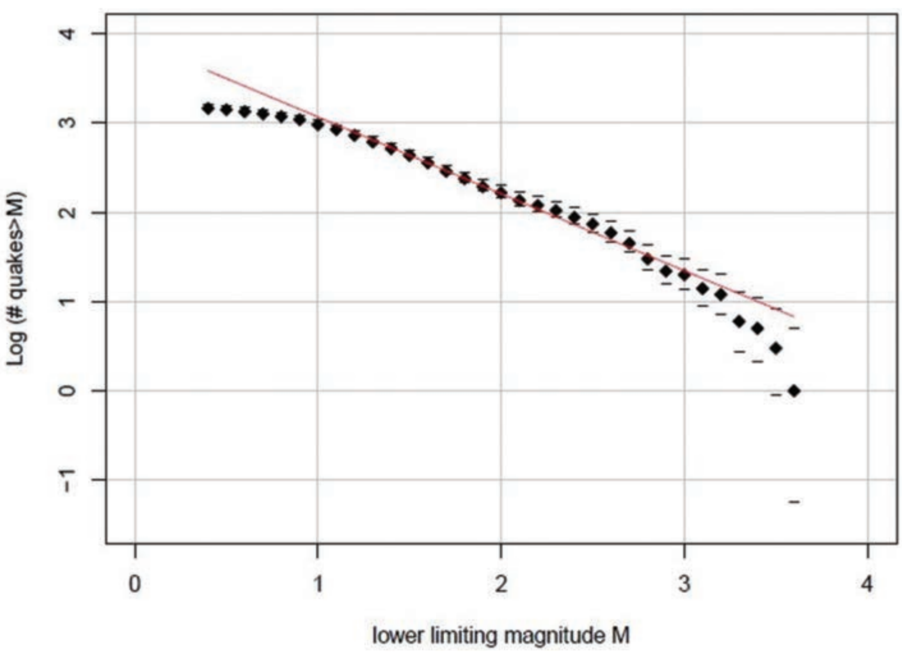

Figure 1 The logarithmic cumulative magnitude distribution of earthquakes for all earthquakes in the set. The red line is a linear function with a slope of -0.9 similar to values reported elsewhere [11]. 95% confidence intervals are indicated under the assumption that the underlying process obeys Poisson statistics.

Self-organizing behaviours can also be found on other scales in nature, and they are not limited to the animal kingdom either. It only takes a short trip from here to the dunes to see that structures made up of loose grains of sand display certain patterns that do not follow trivially from the properties of those grains of sand. As the slope of a dune increases, the chance of larger and smaller cascades and streams of sand increases. These sand displacements make the dune slope smaller. There is thus a feedback mechanism that ensures that dunes retain a certain shape and remain permanently in what is called a *self-organized critical state* [3].

Another example of this are earthquakes. A geophysical rock stratum can be unstable from fractures, such as where continental plates meet and from other natural causes, but fractures come in all sizes; from hundreds of kilometres to a few micrometres, there is a very broad spectrum. Those very small fractures will be described as brittle rocks rather than fractures, but they are actually part of a continuum [17]. If the pressure on that rock changes, for example because of shifting overlying layers of the earth, or because an internal pressure changes, such as when gas is extracted, it could mean that that rock structure suddenly changes violently, resulting in an expanding wave movement (an earthquake). Because the faults are present over a wide spectrum of sizes, the amount of energy in each of those quakes can also vary over a wide spectrum. This can also be visualized with the aid of diagrams such as Gutenberg–Richter plots [22,25], see Figure 1. What you see here are counts of quakes of a certain magnitude on the Richter scale, in this case for the province of Groningen. Larger quakes are rarer than small quakes, so the counts decrease as the quake strength increases, more towards the right in the diagram. Also very characteristic of complex systems is that in diagrams in which the number of times something occurs against the magnitude of that something, or its energy, or some other property, there seems to be a power-law relationship. If both axes are made logarithmic, as shown here, the result is something similar to a straight line. Of course, there is much more to say about the statistics of the earthquakes in Groningen, but then I digress too much (see [31,32,33] for some relevant websites).

I can not resist the temptation to give you an example of things on an even larger scale. What you see in Figure 2 below are large collections of stars [20]. The photo on the left is a galaxy, number 74 from the catalogue originally prepared by a French astronomer, Charles Messier, who lived in the late eighteenth century, early nineteenth century. This photo of the object is of a much more recent date than the man. The photo on the right is object M3 from the same catalogue, and these types of objects are called globular clusters: also a gravitationally bound collection of stars but much smaller than a galaxy and with a different history and structure. M74 is roughly comparable in size to our own galaxy. Our galaxy, and also M74, has quite a few satellites such as M3. What I want to do next is something that will appeal to all Statistics Netherlands staff among you: a census of the star population. I graphically represent the result of that census in the form that astronomers will recognize as a Hertzsprung–Russell diagram. The picture below shows the result for M3 [28]. The vertical axis shows intrinsic brightness, on the horizontal axis is how astronomers measure colour which is a measure of a surface temperature of the star, with hotter stars more to the left in the diagram. Every point in the diagram is a star, and you will probably notice that points are not evenly distributed throughout the diagram.

The greatest density of dots is distributed as a kind of ribbon that runs from the bottom centre of the term 'Main Sequence' up to where it says 'Red giant branch', and then is draped across the diagram to the left. I will not elaborate about the life course of stars, but what I do want to tell you is that within that ribbon, in the order I just used to describe the diagram, stars increase in mass, from less than once that of the sun, to a few dozens of times that of the sun. The counts of the numbers of stars as a function of their mass, the distribution function for star mass, also shows a power law [4]. That says something about the self-regulating processes that determine the mass of stars when they are formed. I'm cheating a bit, because the diagram on the left is not the census of M74. Instead, it's the census of our own Milky Way, very similar to M74, where the GAIA satellite [12] gives us enough precision measurements for roughly 1.3 billion stars. The dot density in the chart is so great in places that a colour scale is used to indicate how many stars are in that location in the chart. Also in galaxies, there is a distribution function of the star mass that behaves like a power law. The differences in these two diagrams, the systematics revealed by this census, shows that the majority of stars in the globular cluster formed almost simultaneously, while galaxies have multiple generations of stars. That is, the star formation process has active periods and also less active periods. This also has consequences for the spatial structure of these two types of systems. This too is an important aspect of complex systems: it may well be that a complex system does not have just one equilibrium, it can instead show periodic phenomena, or a few discrete quasi-equilibria between which it jumps from one to the other at unpredictable times.

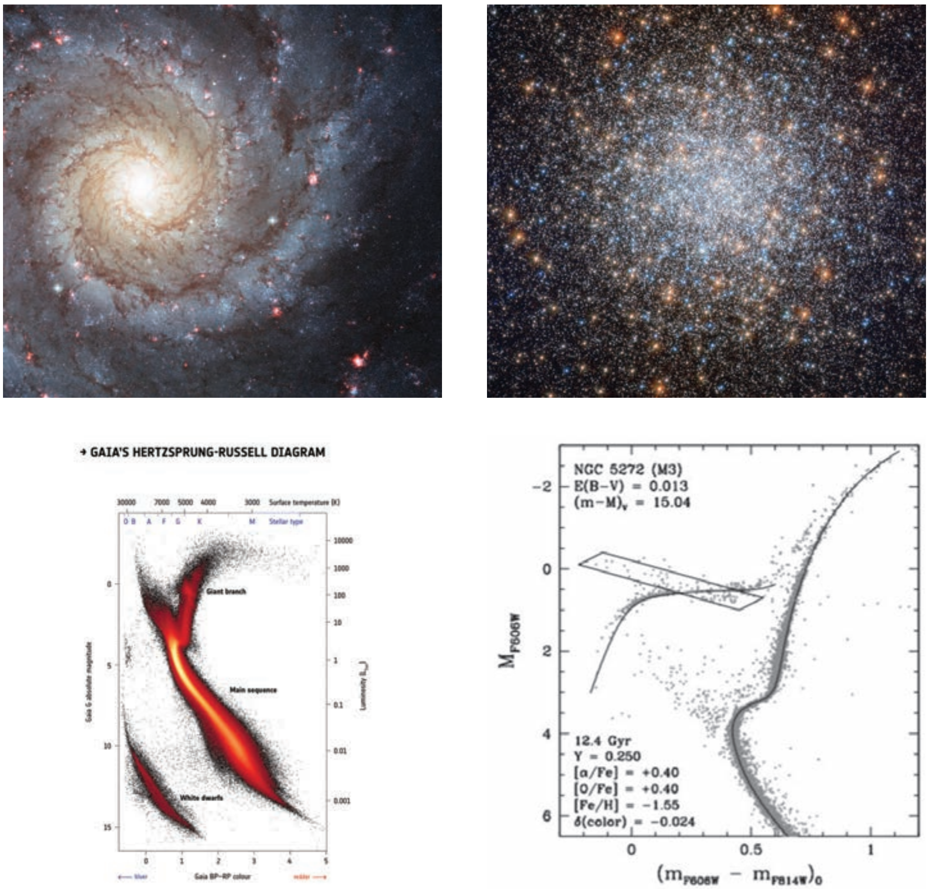

Figure 2 A large collection of stars and the Hertzsprung-Russell diagrams.

After this whirlwind of examples, I now come to the point that I want to take a step back and try to formulate a definition for a complex system. I hope it is clear that such a system consists of many parts that have a certain degree of autonomous existence, such as termites or stars, but also form certain structures in each other's presence such as termite mounds and galaxies. For the purposes of Statistics Netherlands, these autonomous units are people or companies and the structures are the Dutch society and the economy. As a system, they will exhibit certain characteristics such as emergent behaviour and self-organization, and in determining distribution functions of properties, skewed power laws will be the rule rather than the exception. It may also be expected that in the time evolution of such systems there will be sudden transitions interspersed with longer periods of apparently stable equilibria.

A certain amount of abstraction is of course necessary to explore patterns that govern structures in society and the economy. From statistical physics and mathematics, we learn that it is easier to start with homogeneous systems, in which all parts are identical, and only later to investigate the influence of heterogeneity and differentiation. Incidentally, there is also a need to do both, because it helps to assess whether there is a difference between homogeneous and heterogeneous systems, and if so, which is most beneficial, for example, for their long-term stability.

Besides making a distinction between homogeneous and heterogeneous systems, it is also useful to take into account the timescales on which the units persist and how this relates to the timescales of interactions. If units mainly interact with a certain subset of other units over their entire lifespan, the system as a whole is relatively rigid, and it makes sense to use the methods and mathematics that are used for research into (complex) networks. Incidentally, the rigidity of that network structure does not imply anything about the great dynamism that can arise in signals that propagate over that network, such as infections with a virus or a conspiracy theory, or congestion over a road network.

On the other hand, it is also possible that parts of a system have frequent but fleeting contacts. In astronomy and statistical physics, computers are usually used to perform many-particle simulations of such systems. After many repetitions of simulations with slightly different initial conditions, or adding noise, the statistical properties of the system are then mapped out. In computational social sciences, the equivalent is called *agent-based modelling*. Thus, this kind of simulation approach is not primarily to predict the future of a system, but to examine how model parameters, i.e. the properties of particles or agents, determine the behaviour of the entire system.

Sometimes it makes sense to describe a system hierarchically. Detailed modelling of interactions at the microscopic level (human to human, or company to company) is abandoned, and in its place a much smaller system of blocks interacting with each other is modelled. Within each block, in the context of this framework, there are a large number of units that have a certain property in common, and the interaction between the blocks then concerns the average interaction between units belonging to those blocks. This reduction in the number of units in the model will have an effect on the modelling. Experience shows that there is a trade-off in the sense that the mathematical equations, which calculate how a system behaves, will receive more terms and in particular non-linear terms. For example, it can be considered that not only the mean of the population within each block returns as a variable in those equations, but also the variance, the skewness, and the kurtosis of properties within each block, or correlations between properties. I will come back to the importance of non-linearity later.

Since all three situations can occur in interactions between companies, or interactions between people, my research on complex systems will therefore also have different lines, one for each of these approaches: the theory of complex networks, agent based modelling, and dynamical system theory [9, 29, 30].

About solitons, bifurcations, and chaos

So far, I have bravely resisted the temptation to switch from English and speak a language in which it is easier to be precise and unambiguous; I mean the language of equations and formulas: mathematics. I will try to maintain that, but I still need it every now and then since I also want to be able to quantify. For example, what determines the value of the exponent in those power laws of distribution functions? Stepping over a network, how many steps are the minimum and maximum steps required to get from one unit to another? How fast do signals such as infections, or opinions, or innovations, or money travel across a network? As soon as we want to measure these indicators in society or the economy, and compare those measurements with explanatory models, it is necessary to convert the qualitative descriptions into numerical parameters and indicators.

I just mentioned that in the line of dynamical systems research, the mathematical formulation of the model of reality will often consist of a set of coupled differential equations. The description of the system as separate elements with their interactions is often referred to as a *microscopic* approach. Once I start grouping those elements and modelling the interactions as interactions between groups, I have what is called a *mesoscopic* approach. For the sake of completeness, I would like to mention that a *macroscopic* approach is one in which I would put all elements in one block and create equations for a set of properties of that entire set. The physicists among you recognize the macroscopic approach as thermodynamics, where that microscopic approach is the field of statistical physics. In the applications to Dutch society the analogy is somewhat flawed, because the systems of thermodynamics typically have of the order of 10^{24} elements, while Dutch society has only roughly $17.5 \cdot 10^6$ persons.

Each group, each block in the dynamic system, is fully described in this approach by a number of group properties, these are the dependent variables $y_1, y_2, y_3, y_4, \ldots, y_N$ in a model. For the sake of convenience, I will assume that these variables can take not only discrete values, but all values in an interval. Those properties can differ from block to block and can change over time. A change in time of a dependent variable is indicated as a rate or time derivative. Each of those time derivatives can be related to that set of variables themselves, or products thereof, or powers, or other non-linear combinations. It is even possible that local spatial gradients play a role in the development of time and therefore also derivatives with respect to a position x occur. For the sake of convenience, I put all those dependent variables into one vector $Y = (y_1, y_2, y_3, y_4, \ldots, y_N)$, and summarize all those possible combinations of products, derivatives and so on, with $F(Y)$, so that the system of coupled non-linear differential equations takes the following innocent-looking shape:

$$\frac{\partial Y}{\partial t} = F(Y).$$

In practice, this can take on a very complicated form, so it is not useful to specify this in more detail now. I just want to show you a few examples of what we can possibly expect to see happen. I think it is appropriate to show a specific equation whose solution shows something interesting, and which will be very familiar to my colleagues from the University of Amsterdam in particular:

$$\frac{\partial y}{\partial t} + \frac{\partial^3 y}{\partial x^3} - 6y\frac{\partial y}{\partial x} = 0.$$

This is the Korteweg–de Vries equation, for one dependent variable y, which depends on time t and a spatial coordinate x. There is a lot to say about this equation, but that really goes too far afield. The point here is that it is an example of the kind of equations, with nonlinear terms, that we will also distil from descriptions of interactions between groups of people, or between groups of companies. The interesting solutions of this equation are running waves over the coordinate x. These are generated spontaneously in the system, even from a smooth initial situation, which a linear system will never do. It is also remarkable that two different wave solutions, which will have a different speed, can collide and emerge unscathed from that collision.

This is nothing special for linear equations, but very unexpected for equations with non-linear terms. These types of traveling wave solutions produced spontaneously in the system are referred to by the term *soliton*. A nice detail is that the real-world phenomenon was actually accurately described by a Scottish marine engineer John Scott Russell, some sixty years before the mathematical equation and analysis was published. I would like to remind anyone who is wondering how this could be of importance to the field of activity of Statistics Netherlands that there are sufficient examples, even in recent history, of social or ideological groups and political parties in the Netherlands, which at first sight appear to be completely homogeneous and yet spontaneously undergo a schism and move apart in subgroups. At the very least, it is interesting to research the circumstances or characteristics of groups that influence the likelihood of schisms and social polarization.

For a long time, it was common practice in both physics and economics to describe systems as equilibria and to approach changes over time as reversible processes. It is partly under the influence of Prof. Ilya R. Prigogine (1917–2003, Nobel Prize for Chemistry 1977) that research into dissipative systems and systems that are far out of equilibrium has boomed. A beautiful and very simple example of the latter is the sudden formation of crystals in super-cooled water. Under the right conditions, it is possible to cool water to well below 0 Celsius, while still remaining liquid. A very minimal disturbance can suddenly trigger a transition in that system, and crystal formation occurs on a very short time scale. For this behaviour, too, an equation can be formulated that describes the time evolution of the relative proportions of water and ice crystals, which indeed includes non-linear terms. Much the same formalism has also been used to describe traffic congestion, where the two phases, water and ice, are replaced by moving and stationary cars. We have probably all experienced sitting impatiently in a queue of stationary cars and finding that suddenly everything starts moving without any indications of what caused the traffic jam in the first place. Both are a manifestation of self-organizing behaviour. In 1972 Herman, Lam and Prigogine [13] showed that the statistics of journey times over certain routes in the US, UK and Australia, and the amount of time that cars had to stand still on those routes, could be well explained by this model. The Netherlands not only has a high population density, and therefore car density; logistics and freight transport by road play an important role in the Dutch economy. It is therefore good for policy-making and planning in the field of road construction and the environment to know which parameters most effectively favourably influence the flow of traffic. It is therefore obvious that the values of precisely those parameters for traffic behaviour in the Netherlands should be monitored.

In the examples I discussed so far, the complex systems spontaneously showed certain unexpected behaviour as a function of time, while the model parameters of those systems retained certain fixed values. If in a model for a complex system the parameter values are varied, the solutions of the model naturally change, but in many cases such changes can be regarded as gradual, continuous, deformations. In some cases, however, something else happens. The example I want to quote here is the so-called logistic map (see Figure 3). It is sometimes also referred to as the predator-prey equation since it describes the precarious balance between these two in a closed system. This again describes the time evolution of a system with a dependent variable x, but in discrete steps instead of as a continuous function of time:

$$x_{t+1} = r x_t (1 - x_t).$$

The single parameter r controls the behaviour of this system. When r is less than a certain limit, x tends to a single equilibrium value. However, as soon as that limiting value for r is exceeded, the behaviour of the system changes abruptly. Instead of converging to a single equilibrium value, the x jumps back and forth between two different values. At a second, somewhat larger limiting value, the system again abruptly changes its behaviour and jumps between four different values. These abrupt doublings of the number of values between which the system jumps keeps repeating at successive limiting values, up to a certain point. After this there are values for r for which x no longer shows any periodic behaviour. A diagram can be made of the equilibrium values obtained as a function of the value of the parameter r. Every splitting or fork in this graph is called a *bifurcation*. Mathematicians distinguish different types of bifurcations. For the larger values of r where random values for the dependent variable x succeed each other, they fill up the entire range of possible values. The exact order of values for x is extremely sensitive to the initial value for x. In mathematical terms this is called a *chaotic* system. There are macroeconomic models in which inflation and labour costs (or unemployment) are modelled in a similar way. The validity of such models is controversial, so I will not discuss them in detail. However, the behaviour of model systems that bifurcate or become chaotic for certain combinations of parameters is known and well researched for many different systems, including the climate of our planet, and is certainly of importance for understanding our economy.

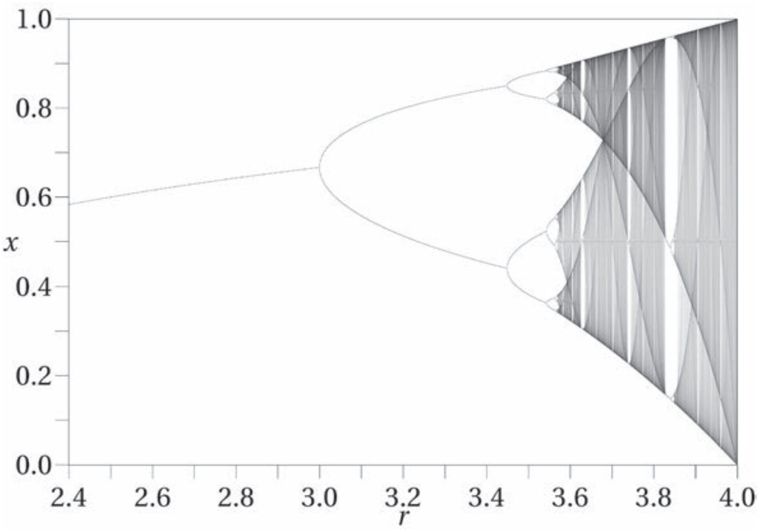

Figure 3 Logistic map.

One reason for this importance is fairly obvious. If an accepted economic model can display this type of behaviour, it is of course important to determine, from measurements, what exactly the value of the crucial parameters is, because planning bureaus need the correct values of that parameter, for example. In order to be able to make a good, accurate measurement, researchers at Statistics Netherlands must of course know how the model works, so that the correct data can be collected as input for unbiased statistical estimators of those parameters. The second reason is less obvious. Statistics Netherlands publishes many indicators, including economic indicators, on a monthly or annual basis, that is, at discrete times. Not only the planning offices, but also other parties that themselves play an active role in the economy, use these data to make estimates or predictions themselves. This means that decisions with economic consequences are also based on the time series published with a certain cadence. The example of the logistic map shows that if an unfortunate choice is made of the weight of the existing data in model predictions and the subsequent actions, the indicator in question can start to display chaotic, unpredictable behaviour with potentially disastrous economic consequences. This picture is less far-fetched than it seems.

Stock trading based on computer algorithms became fashionable in the 1990s. Since then, intra-day trading has evolved into high frequency trading where, for specialists in this field, even to have nanoseconds faster algorithms than competitors can generate enormous financial profits. In the early days of these algorithmic trading systems, there were not enough security measures in place to limit the strength of possible feedback. It is likely that this contributed to the 2010 flash crash [16], and it may also have caused more variability ('volatility') earlier on (the role of expectations in boundedly rational agents in the economy is discussed extensively in [14]). There is no unambiguous evidence that this still significantly affects the stock markets. Even though predictions for future developments in the economy fall outside the scope of Statistics Netherlands itself, it is important for public administration to be familiar with, and to be prepared for, such feedback effects.

Fractals everywhere?

In the course of my lesson, I have shown that complex systems that occur in nature have certain characteristic features that can be captured in, i.e. reproduced by, mathematical models that fall under the collective term non-linear dynamical systems. I haven't yet paid much attention to complex networks, so I'll get to that in a minute. As a bridge between the two, I would like to add another mathematical concept, namely that of the dimension of objects, or of spaces. I probably won't have to tell you that a point without dimensions has dimension 0, and a line has dimension 1, a plane or something in it such as a circle has dimension 2, and so on. I will not tease you with 4-dimensional spacetime and with curvature tensors. I do want to discuss some remarkable objects, which can be constructed recursively, for example. Recursive means: by repeating a certain simple recipe of steps infinitely often. Think for example of the Sierpinski carpet and the Menger–Sierpinski sponge. A tastier real-life example is Romanesco broccoli. There are many more variations of these types of objects associated with the names of famous mathematicians like Peano, Cantor, or Mandelbrot. There are several ways to precisely determine the dimension of an object mathematically. The Hausdorff dimension is a well-known way, and that dimension is neatly integer for the more ordinary objects, but not for the Sierpinski carpet and Menger–Sierpinski sponge. The Hausdorff dimension is fractional, which is where the name *fractal* for these types of objects originates. To make a connection with non-linear dynamics, think of the Lorenz butterfly, which is the path in a three-dimensional space that the solution of a non-linear dynamic system with three dependent variables over time follows. Obviously, that path isn't perfectly periodic, but it doesn't fill the whole space either. It is not uncommon for nonlinear dynamical systems to do such a thing, they have what is technically called a strange attractor and strange attractors have the property of having fractional dimensions. Within that strange attractor, that path does fill that particular sub-space and thus has the characteristics of chaos, which is also called *low-dimensional chaos*.

That brings me to the concept of a network. For the typical purposes of Statistics Netherlands, each circle can represent a person or, for a different purpose, a company. The lines connect two circles if those two people also know each other. Statistics Netherlands has built up a network based on people who are related to each other according to the basic register of persons (BRP): parents, children, brothers and sisters, et cetera. In addition, a network can also be built on the basis of people being part of the same household, and on the basis of being neighbours, and on the basis of going to the same school or other educational institution, or having the same employer. All those networks are of course connected because they involve the same people: about 17.5 million of them. The data needed to record all these relationships had been available to Statistics Netherlands for a long time, but it had never been stored in this combined form before. There are of course also other ways in which people could know each other that Statistics Netherlands cannot detect. Friends and acquaintances on social media platforms, or from a neighbourhood cafe, gym or music or hobby association are not stored in the Statistics Netherlands' databases. Nor does Statistics Netherlands know that there are people you see every morning when you are waiting for that train to work in the early morning, and who, like you, are preparing for the day with a cup of coffee.

It is evident that not every resident of the Netherlands knows every other resident in this way, so far fewer lines run through that diagram than is maximally possible. It is also not the case that everyone knows roughly the same number of people. The distribution function that describes how many people have a relationship with only 1, or 2 others, or with 5, or with 100 or more, can also be described with such a power law that I also showed before for earthquakes. It is of course not for nothing that I have placed this description of networks under the heading fractals, so you probably already guessed, the exponent in that power law is something like a dimension of the network, and that dimension is fractional. This dimension for a network also follows a different definition than the Hausdorff dimension definition I just used. The Hausdorff dimension for networks can be constructed, but will generally not yield exactly the same value. Also, there are some practical limitations in dimension measurements in the sense that those mathematical fractal objects, like the few I just showed, keep repeating themselves to an infinitesimal scale, and in the world around us there is always a smallest scale in space or time at some point.

There is a second problem with correctly estimating that dimension, or other measures of the interconnectedness in the social network of the Netherlands, and the cause of this I have just mentioned. Statistics Netherlands only has data for certain types of relationships and not for all types. That means that the real network of contacts between people is measured incompletely, and moreover that the missing relationships are not random. Selectivity and incomplete data are the worst enemy of official statistics, not only in sample-based surveys, but in all forms of data. An equally serious problem is that the relationships that Statistics Netherlands can establish are not necessarily a good reflection of actual contact between people. As an example, consider a large company with many different business locations. Two people can very well work for one and the same company, but one for a location in Groningen and the other for a location in Vlissingen. In that case it is not evident that these two people know or meet each other. At the very least, this must be taken into account in analyses. The intention is that Statistics Netherlands will publish reliable new statistics that provide information about the statistical properties of relationships between people. An example are statistics on the geographic distance between the place of residence of adults and the place of residence of their elderly and vulnerable relatives, for which they may at some future time have to provide informal care [8]. Not unimportant when publishing this new social-network based statistics is that this must be done under strict conditions of privacy protection, which has therefore also become the subject of Statistics Netherlands research projects.

In the case of relationships between companies, less data are available to Statistics Netherlands than there are for people. That relationship will typically be one in which companies are suppliers or buyers of services or goods from each other. Apart from the accounts of those companies themselves, there is no registration of individual transactions and there is also only very limited reporting, usually in the form of aggregates as annual sales. Those annual turnovers follow a power law [24] and are therefore perhaps an expression of a self-organized critical system. Of course, every transaction between companies also involves money, so banks do have a partial view of corporate networks for their own customers. Discussions are currently underway on whether, how, and in what form statistical analyses by Statistics Netherlands on such banking data could be made possible. Even if that can ever be arranged, there is still a problem of selectivity there, partly because there is a very wide variety of business forms, with all kinds of peculiarities of ownership and administrative structures. There is great importance for public policy and therefore for official statistics to know more about the connections between companies within the Dutch economy. Greater connectivity may mean, for example, that if a particular company ceases to exist for whatever reason, there are other companies that can take over the role and the system as a whole is therefore relatively resilient and the flow of goods and services across that network remains stable. However, it can also mean that the bankruptcy of a particular company causes a whole cascade of bankruptcies through that network of associated companies. More concretely, Statistics Netherlands is currently investigating for a specific regional cluster of companies what dependencies between those companies play a role in facilitating a transition to using sustainable sources of energy [18].

It is certainly not impossible to take into account and correct for bias caused by incomplete data. This is a well-known field of activity for Statistics Netherlands, but the context of network data is new and more research is therefore needed to develop accurate estimators and good imputation methods.

As I just indicated, it is already important for official statistics to determine properties of that set of relationships between persons or between companies. It is therefore a small step to not only want to map the networks themselves, but also to measure the behaviour of dynamic processes on that network. How does the chance of cascades of bankruptcies depend on the connectivity of a business network? How quickly do innovations penetrate such a network? What is the probability that healthier lifestyles are contagious across a social network and how fast does that work? Does that also work with attitudes towards conspiracy theories, or with the spread of infectious diseases?

Now I show a short animation of such a process that takes place on a network. It is only a toy model because it has very simple rules (see the video of this Inaugural Lecture on https://youtu.be/w-SfnnVfu9M). As you can see, spheres change from black to red if they connect to a sphere that is already red. Then the colour of the spheres slowly turns black again. Since this process only occurs when there is a connection, it is no surprise that the amount of connections will play a role in the speed at which red dots appear. But not only the average amount of connections plays a role. If there is a cluster of points that are interconnected, but have no connection to the outside, there is no mechanism to get red dots within that cluster. Very subtle details in a network of contacts can therefore have major consequences for distribution models. It is therefore important, for public safety, for public health, and for a robust economy, to know exactly which subtle properties of a network must be mapped by Statistics Netherlands. Only then can the planning offices be enabled to correctly calculate the effect of policy measures. In this context, policy measures are either the kind of measures that change the connectivity of the network, or measures that send a signal across the existing network: such as changing the colour of the dots. It is not at all self-evident that researchers at Statistics Netherlands already know what is important to measure in terms of Dutch society and the economy. That is precisely why knowledge about how models are constructed must be brought to Statistics Netherlands.

Complex networks are a good translation into mathematical objects and rules of how people or companies can form lasting relationships. In its most extreme form, all connections in the network are fixed, and research revolves around processes and signals that move through the system over those network connections. That is probably not a very realistic approach to reality. Not all contacts between people or companies are long-lasting. It is therefore desirable to be able to simulate more flexible systems, in which the timescale on which processes run through a system and the timescale on which connections appear or disappear in the network are more comparable. This requires different mathematical approaches than when those time scales differ greatly, and thus also influences what Statistics Netherlands should measure in terms of society, when we translate those mathematical insights back into the field of official statistics.

That brings me to that third line of research, the particle simulations or agent-based models. Astrophysicists among others, have long resorted to the use of simulations, in which the particles are, for example, entire galaxies and the system the large-scale structure of the universe [26]. There are several reasons for resorting to this approach. An important reason is when little is known about the typical timescales of various interactions between the agents, or when those timescales are close to each other. In this case, it is more objective to specify the rules for interaction, as well as the rules necessary for the system to evolve one step ahead in time. It is then left to computers to perform a new simulation very often, with slightly different initial conditions each time, so that one can draw statistical conclusions from the ensemble of all those simulations. To investigate the influence of parameters of those rules for interactions, you repeat that process for a series of values of those parameters. You will understand that this is very computationally intensive, and progress in this area also depends on both faster computers and smarter computing algorithms. It is now within reach to conduct this type of research for models with tens of millions of agents at the same time. For me, the aim is to investigate the mechanisms underlying emergent social phenomena such as segregation and inequality with these tools.

Measurement data and model-building

Measurement data as a means of excluding models — model-building to determine what needs to be measured. It is often said, and inappropriately on more than a few of those occasions, that society or the economy has become more complex over the past century. Sometimes the only thing that is meant is that there have been changes and that old categorizations of the work people do, or their life course, no longer fit well with modern society and economy. An example of change in a system that genuinely can lead to complexity is through specialization. When a company grows, individual sectors of that company can become increasingly separated. These individual units can even separate completely and then independently enter into associations with other companies. So that means that in the entire economic system more autonomous units have appeared with interactions between some, but not all units. Successively mechanization, automation and computerization of production processes have indeed resulted in much more specialization over the past century. The concept of production chain is still used within Statistics Netherlands to indicate that there are a number of steps between extracting raw materials and selling an end product to a consumer. Perhaps it is now better to speak of production networks, or production paths within the enterprise network.

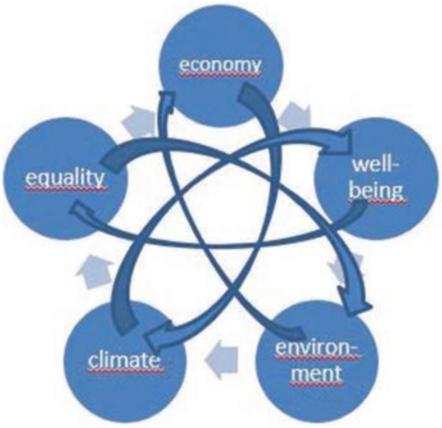

Figure 4 Sustainable development goals.

In other cases, the system has always been complex but its widespread recognition originated more recently. An example of this are the sustainable development goals (SDGs), which were drawn up on behalf of the United Nations in response to a report by famous economists, Stiglitz, Sen and Fitoussi [27]. See Figure 4. In the diagram I highlight a few areas by way of illustration, with which I certainly do not want to suggest that other goals in the SDGs are less important. What I want to point out is that economic activity has an impact on people's well-being, both favourable and unfavourable, on the living environment and the climate, and also on equality between people. Each of these also has an impact on each other and on economic activity. In recent years, economists have spoken of a worldwide slowdown in (production) growth [21]. If, for example, we look at easily measurable indicators such as the change in gross national products or such as inflation in most countries, this does indeed seem justified. Since this is the collective economy of the whole world, that slowdown is something that is caused within that system itself. If the interaction between all elements in this system were only linear, this would not occur, so limitation to growth is quite direct evidence for non-linear effects between those areas mentioned in the SDGs. Feedback is also built into the entire system, in various ways. An example of direct feedback is that between equality and the well-being of persons. The skewed distributions of income and wealth, even in the most egalitarian countries, is at the very least an indication that socio-economic systems are in a self-organized critical state.

At the level of the United Nations, there is thus a recognition that these interactions are at least as important as the subjects themselves. Within the Netherlands, attention is already being paid to this at public policy level by having the 'monitor brede welvaart' [6] drawn up by Statistics Netherlands, in collaboration with other partners, and discussing it in a separate session in parliament. This monitor is certainly valuable and in my opinion one of the most important modern publications of Statistics Netherlands. However, out of necessity, it still pays mainly attention to units, rather than to the interactions between them. There are various models used by economists to explain correlations in various time series shown in that monitor. These models therefore presuppose parameterized mechanisms, and these should be tested. There may already be data of sufficient quality to rule out the validity of some of those models. In order to do this with sufficient expertise, it is essential that knowledge about modelling complex systems also comes to life within Statistics Netherlands, so that it becomes clear which model predictions are sufficiently distinctive to have the option of excluding models by means of statistical tests.

Another important alternative route is to test directly from time series whether correlation could also imply causation. In fact, the reverse is actually more feasible: can a hypothetical causation mechanism be falsified using cross-correlation data? In the context of official statistics, I have taken a first small step [7], but there is much more notable work [15]. I expect a lot from collaboration with UvA colleagues who have a lot of expertise in the field of testing causation.

I am almost ready to complete this public lesson, but there is one more concept that is important to enable building that bridge between measuring and knowing, namely ergodicity. I avoid spelling out the formal definition [19] (for an informal but somewhat more extensive description, see, e.g.,[23]), but when a system is ergodic it means that the average of a time series of a system indicator — something that is measurable from the system — yields the same as when I take the average of that indicator over many realizations at a single point in time from exactly the same system. Although there are many systems in nature that meet this requirement, it is not a self-evident characteristic. For example, in the research line of agent-based modelling, the aim is to compare the average of certain indicators over many realizations of a model system with averages of time series of the same indicators. This is only valid if we assume, or preferably can prove, that the system is ergodic. A striking example of a non-ergodic system is that of random multiplicative growth [2]: a gambling game in which, dependent on many repeated throws of a fair coin, at each toss you gain 50% of all your money or lose 40% of it. Ensemble averaging might suggest this is a sure win for the gambler, but time averaging of single realisations suggests otherwise.

The statement of the physicist Heike Kamerlingh Onnes from 1882 "Through measurement to knowledge" is evidently also applicable to society. My research plan therefore also fits in with, or is an extension of, Tinbergen's tradition. National statistical institutes such as Statistics Netherlands measure and report on society according to common standards with as much the same population demarcation as possible. The next step is to take new measurements to understand not only what society *looks like*, but also how society *works*. Ultimately, it is the task of Statistics Netherlands to support public policy with facts. If it is possible for Statistics Netherlands to demonstrate how the economic and social system works, or at least that certain assumptions on how it operates are not correct, this can be crucial for effective and meaningful governance. Via the SDGs, we now realize that society and the economy are complex systems. That makes the knowledge concerning complex models for Statistics Netherlands 'need to know' instead of 'nice to know'.

References

1 D. Adams, *The Ultimate Hitchhiker's Guide*, Wings Books, 1996, pp. 90–91.

2 A. Adamou, What is ergodicity? https://youtu.be/VCb2AMN87cg, 12 February 2021.

3 P. Bak, *How Nature Works: The Science of Self-organized Criticality*, Springer, 1996.

4 I.K. Baldry, K. Glazebrook, S.P. Driver, On the galaxy stellar mass function, the mass–metallicity relation and the implied baryonic mass function, *Monthly Notices RAS* 388 (2008), 945–959.

5 A. van den Bogaard, Het CBP, wiskunde en praktijk in wording, *Nieuw Archief voor Wiskunde* 5/1 (2000), 294–300.

6 CBS, *Monitor Brede Welvaart & de Sustainable Development Goals 2020*, cbs.nl.

7 G.E. Comi, R.J. Fitzner, S. Kolumban, F.P. Pijpers, R.M. Pires da Silva Castro, R.A.J. Post and A.J. Vromans, Causal effects of government decisions on earthquakes in Groningen, *Proceedings of the Study Group Mathematics with Industry, 2018* (2019), swi-wiskunde.nl.

8 M. Das and E. de Jonge, Zelfredzaamheid van ouderen en gebruik van Wmo (2020), cbs.nl.

9 R.L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Benjamin/Cummings, 1986.

10 C.M. Dobson, Protein folding and misfolding, *Nature* 426 (2003), 884–890.

11 B. Dost, F. Goutbeek, T. Eck, and D. Kraaijpoel, *Monitoring induced seismicity in the North of the Netherlands: Status report 2010*, 2012.

12 Gaia Data Processing and Analysis Consortium (DPAC), https://www.cosmos.esa.int/web/gaia; Carine Babusiaux, https://wwwhip.obspm.fr/~carine.

13 R. Herman, T. Lam and I. Prigogine, Multilane vehicular traffic and adaptive human behavior, *Science* 179(4076) (1972), 918–920.

14 C. Hommes, *Behavioral Rationality and Heterogeneous Expectations in Complex Economic Systems*, Cambridge University Press, 2013.

15 G.W. Imbens and D.B. Rubin, *Causal Inference for Statistics, Social and Biomedical sciences: An Introduction*, Cambridge University Press, 2015.

16 A.A. Kirilenko, A.S. Kyle, M. Samadi and T. Tuzun, The flash crash: high-frequency trading in an electronic market, *Journal of Finance*, (2017), forthcoming.

17 I. Main, Earthquakes as critical phenomena: Implications for probabilistic seismic hazard analysis, *Bull. Seismological Soc. Am.* 85 (1995), 1299–1308.

18 C.E.S. Mattsson, F.W. Takes, E.M. Heemskerk, C. Diks, G. Buiten, A. Faber and P.M.A. Sloot, Functional structure in production networks, *Front. Big Data* (2021).

19 W.D. McComb, *The Physics of Fluid Turbulence*, Clarendon, Oxford, 1990, p. 39.

20 NASA, nasa.gov; ESA, esa.int; The Hubble Heritage, hubblesite.org.

21 OECD Interim Economic Outlook, 2019, https://www.oecd.org/economy/outlook/interim-economic-outlook-march-2019.

22 F.P. Pijpers and V. van Straalen, Trend changes in tremor rates in Groningen, CBS Technical Report (2018), https://www.researchgate.net/publication/330797660.

23 F.P. Pijpers, Five sessions on turbulence (1993), p.10, https://www.researchgate.net/publication/237581928.

24 F.P. Pijpers, Applications of complexity theory in official statistics, CBS Discussion Paper (2018), https://www.researchgate.net/publication/322581886.

25 D. Sijacic, F. Pijpers, M. Nepveu and K. van Thienen-Visser, Statistical evidence on the effect of production changes on induced seismicity, *Netherlands Journal of Geosciences* 96 (2017), 27–38.

26 V. Springel et al., Simulations of the formation, evolution and clustering of galaxies and quasars, *Nature* 435 (2005), 629–636.

27 J. Stiglitz, A. Sen and J.P. Fitoussi, Report of the Commission on the Measurement of Economic Performance and Social Progress (CMEPSP) (2009), https://www.researchgate.net/publication/258260767.

28 D.A. VandenBerg, P.A. Denissenkov and M. Catelan, Constraints on the distance moduli, helium and metal abundances, and ages of globular clusters from their RR Lyrae and non-variable horizontal-branch stars. I. M3, M15, and M92', ApJ. 827(2) (2016).

29 F. Verhulst, *Nietlineaire Differentiaalvergelijkingen en Dynamische Systemen*, Epsilon Uitgaven deel 4, 1985.

30 G.M. Zaslavsky, *Physics of Chaos in Hamiltonian Systems*, Imperial College Press, 1998.

31 https://cbs.nl/dossier/ground-movements-in-groningen.

32 https://knmi.nl/nederland-nu/seismologie.

33 https://sodm.nl/sectoren/gaswinning-groningen.