What Is a Spatio-Temporal Model Good For?

Validity as a Function of Purpose and the Questions Answered by a Model

Simon Scheider[1]
Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Judith A. Verstegen
Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Abstract

The concept of validity is a cornerstone of science. Given this central role, it is somewhat surprising to find that validity remains a rather obscure concept. Unfortunately, the term is often reduced to a matter of ground truth data, seemingly because we fail to come to grips with it. In this paper, instead, we take a purpose-based approach to the validity of spatio-temporal models. We argue that a model application is valid only if the model delivers an answer to a particular spatio-temporal question specifying some experiment including spatio-temporal controls and measures. Such questions constitute the information purposes of models, forming an intermediate layer in a pragmatic knowledge pyramid with corresponding levels of validity. We introduce a corresponding question-based grammar that allows us to formally distinguish among contemporary inference, prediction, retrodiction, projection, and retrojection models. We apply the grammar to corresponding examples and discuss the possibilities for validating such models as a means to a given end.

2012 ACM Subject Classification Computing methodologies → Discourse, dialogue and pragmatics

Keywords and phrases validity, fitness-for-purpose, spatio-temporal modeling, pragmatics, question grammar

Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.7

Supplementary Material *Dataset (Source Code)*: `https://github.com/simonscheider/ModelQuestions` [35], archived at `swh:1:dir:c63d8dc4808114f58911f1870afaa462b6338d4f`

Funding *Simon Scheider*: Supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 803498).

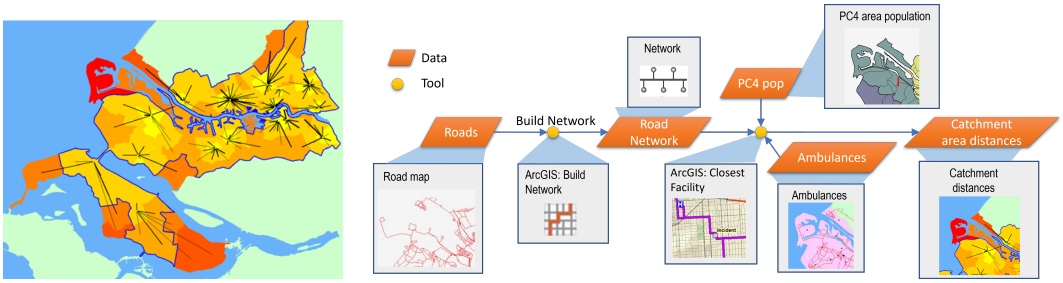
1 Introduction

The concept of validity is a cornerstone of science. Valid methods make science "scientific", while invalid methods are bound to deliver falsehoods, i.e., invalid statements. Given this central role in science, it is somewhat surprising to find that validity remains a rather obscure concept, and no less so in geographic information science (GIS) and spatio-temporal modelling. Take e.g. the following definition of a valid test:

> *A test is valid if it measures what it is supposed to measure (construct validity)*

This definition of validity, originally proposed by Messick in the context of educational performance tests [23, 6], has been adopted in software engineering and statistics to talk about the validity of experiments [12]. It sounds reasonable at first, yet it leaves us in the dark about *who is supposing what and for what purpose*. As long as this is unclear, the definition seems to have a circular character.

[1] corresponding author: s.scheider@uu.nl



(a) Accessibility of postcode areas by ambulances in Rotterdam. Red: low accessibility. Yellow: high accessibility.

(b) Workflow for accessibility measured in terms of travel distance over street network, using closest facility analysis and other ArcGIS tools.

Figure 1 Example of a GIS analysis, which corresponds to a transformation of maps.

The fact that validation is *subject to purpose* has repeatedly been stressed by various authors in ecological modelling since the 70s [33, 3, 47]. Yet it has remained unclear to date how the practical notion of *fitness-for-purpose* relates to the theoretical concept of validity. Some authors specified overarching purposes of models and related types of validity, e.g., Caswell [3] states that "both the means and ends of what we call validation" are different between models aimed at prediction and models aimed at insight (understanding) [3]. Yet, a given model can be used for various purposes, and so it might be more adequate to say that *studies* (model applications) have purposes, while models are rather meant to perform a study. In any case, we seem to lack a theory that makes this connection between study purpose and model validity explicit. In data science, this kind of reasoning is often *avoided* by making use of what is called "ground truth data" [48]. Yet, this does not solve the issue either: In practice, ground truth data is often not available, or first needs to be generated by relying on the know-how of experts [40], leading us back to the very problem that we started with, namely how *they* determine validity.

An example may help us illustrate this challenge. How do we know that the map produced by the GIS workflow in Figure 1 is valid, i.e., measures what it is supposed to measure? Note that the map resulting from this workflow is useful because it gives a *valid answer* to a particular question, namely "What would be the accessibility of postcode areas for ambulances in Rotterdam when closing the bridge over Haringvliet?". The answer requires procedural know-how to generate a workflow as in Fig 1b: generating a network from a roads data set, then applying the "Closest Facility" method of ArcGIS, e.g., to a layer of ambulance stations and postcode areas, and then determining catchment area distances for each postcode area. Yet, reducing validity to exposing our model to the scrutiny of data does not work in this case. One of the reasons is that the indicated bridge has never been closed. So validity is dependent on valid inference within a fictive scenario. Furthermore, the data involved already presupposes, and therefore cannot constitute validity: Using an incomplete ambulance dataset will result in wrong accessibility assessments.

In this paper, instead, we suggest that the application of a model is valid because it *answers a particular question* (specifying its purpose), in a way that *successfully substitutes* a particular *spatio-temporal experiment*. Validating a model for a specific purpose, in a nutshell, therefore requires knowing what it is good for, i.e., what sort of experimental question it can answer. Yet, we currently lack any methods for reasoning over such modeling purposes [40] in the validation procedure. To turn this idea into a methodology, we first introduce a *pragmatic account* of information validity based on *questions about spatio-temporal experiments*. We

then propose a meta-model for modelling purposes in the form of a *question grammar*, which distinguishes different forms of model validity for *contemporary, predictive, retrodictive, projective and retrojective* questions. We then show how this meta-model can be used to validate a range of well-known modelling examples.

2 A Pragmatic Account of Spatio-Temporal Model Validity

To gain a better understanding of the validity of models, it is useful to switch perspectives on what we regard as knowledge and what we consider to be information and data (see Figure 2a). Traditionally, knowledge was seen as *declarative*, i.e., knowledge of facts or "*knowing that* something is the case". Furthermore, in modern data science, knowledge is considered extricable from data. However a lot of what we know is implicit and thus actually *procedural*, i.e., it is hidden in *knowing how* something can be done. Knowledge in this sense is a disposition to act, as was suggested by language/speech act philosophers in the early 20th century, e.g., Gilbert Ryle [34]. Pragmatic knowledge models are founded on the idea that *actions and their purposes* form the basis of our (trans-subjective) knowledge within a *community of practice* [16]. We base our ideas on the more recent pragmatic knowledge model of Peter Janich [16, 18] (cf. [40]), which serves as a core for the following considerations.

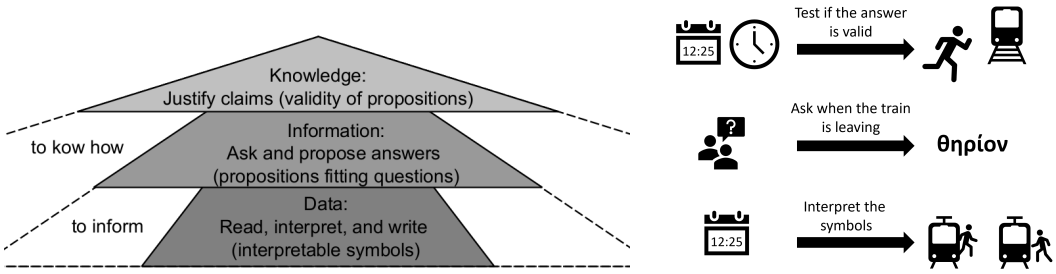
2.1 Pragmatic Knowledge Pyramid

In science and philosophy, knowledge is commonly seen as a collection of true propositions [52]. However, in pragmatics, validity is regarded as a way to establish truth, not the other way around [15]. So, if validity paves the way towards knowledge, then the question arises which actions and purposes are constitutive of validity, and how they relate to data and information [18].

From a pragmatic standpoint (see Figure 2a), symbols are valid data if they are *interpretable* into some shared action schemas (cf. Sect. 2.3.2), and if they can be *read* and *written* with standard syntax (thus easily *shared* on a computer) [18]. In this sense (see Figure 2b), a train departure sign is data because it can be read and written and can be interpreted into a particular type of event that we can all observe. If symbols are not interpretable (because we are unsure what symbols mean) or cannot be read (e.g. a departure sign in a foreign language), they fail to be considered valid on the level of data.

On the next level, *information* is based on specific actions related to the ability to *inform* somebody [17, 18]. The latter means that propositions formed from symbols in fact *answer some question*. Data becomes *valid information* only in case it answers a question since only then it is "informative" for somebody, namely *useful for the purpose expressed by this question*, based on which further action can be taken. In our example, the train arrival sign is in fact information, because it answers a question like "When does the next train to Athens arrive on this platform?", and this is informative because we can act upon the answer, e.g., by arriving at the platform in time.

However, what if the train left earlier and we missed it (Fig. 2b)? To account for such failures, on the next level, we are concerned with validity of *knowledge*. Following Janich [15], a proposition (as an answer to a question) becomes knowledge only in case it is also a *valid claim*, which is a claim that can be justified. *Justifications* can be done in basically two ways, namely by *experiments* or by *inference* from other valid knowledge [16]. The latter is in line with Botkin, as cited in [33], who sees validation as drawing a valid conclusion from arguments. For example, in our example, we can find out whether the arrival board shows the right time by inferring its arrival time from previous locations of the train, or simply by waiting for the train to arrive (an observation experiment).



(a) Pragmatic knowledge pyramid. **(b)** Example: missing a train in Greece.

Figure 2 In the pragmatic knowledge pyramid, data, information and knowledge are the results of actions on different levels. Data consists of symbols that can be read, interpreted and written, information consists of propositions (formed by symbols) that answer a question, and knowledge consists of claims (made using propositions) which can be justified by experiments or by inference.

Note that the pyramid has gaps towards its basis (Fig. 2a): For one, not all knowledge consists of information that informs a person asking a question. This gap indicates that most knowledge is implicit know-how [34], thus even though it is testable and valid, it might have never been used to answer a question. For example, once the train stops, the location of the train entrance is something we know intuitively, without using any kind of explicit information. Likewise, most information that we provide for ourselves is implicit, thus never turned into any data. For example, most of our conversation on this train is never written down.

2.2 Validity of Propositions

The important bit to understand is that *validity* as a general pragmatic concept[2] depends on exactly those action possibilities that define the pyramid. That is, validity can be established by making sure the action schemas that constitute each level can be reproduced (are *trans-subjective*, see [16]) by anyone within a community of researchers: Thus, valid data needs to be readable, interpretable and writable, and valid information needs to be usable for answering a question. Valid knowledge needs to be justifiable, i.e., testable in an experiment or derivable by inference. Furthermore, in case knowledge is explicit by making use of symbols, lower-level action schemas are always implied by default. In this case, *each kind of validity on some level also presupposes the validity of lower levels.* Thus, valid knowledge also requires valid information, as well as valid data. Furthermore, the way a claim can fail (how its proposition can become invalid) can be studied exactly by studying the possibilities for failing actions within this pyramid:

1. A proposition becomes invalid on the data level *whenever symbols used to formulate it are unknown, blurred or un-interpretable.*
2. A proposition becomes invalid on the information level in case *it does not fit a given information purpose by answering the corresponding question.*
3. A proposition becomes invalid on the knowledge level in case it is *not justifiable*, which means either
   a. performances of corresponding *experiments fail*,
   b. or *inferences* (of propositions from other knowledge) *fail.*

[2] Janich uses the term “Geltung”, i.e., validity of claims [16].

This gives us a way to define validity more generally [15], but always relative to some information purpose defined in terms of a question. Certain aspects of this concept of validity are reflected in other work, yet, not in its entirely: Rykiel [33] follows the same distinction between validation of models at the information and knowledge level, naming the former validation and the latter hypothesis testing. Aumann [1] also distinguished the mentioned types of justification on the knowledge level, calling justifications by inference *structural validity*, and justifications by experiment *replicative validity*.

2.3 Knowledge of Spatio-Temporal Experiments

We argued above that validity depends on certain kinds of actions, in particular, the ability to reason with and to perform *experiments* according to a *purpose*. In this section, we outline a pragmatic notion of experimental knowledge underlying spatio-temporal modeling. We start by a recap of Janich's [16] action model (cf. [40]).

2.3.1 Purpose, Success and Failure of Experimental Actions

Let us call an *action* an event that can be *mutually attributed* among members of a community as a person's *responsibility*. Thus: slipping and falling are not actions, but using a slide in a swimming pool is[3]. Knowledge is grounded in *schemas* which are capacities to act that are shared within this community and which can be *actualized* in actions or artefacts (Fig. 3). We call the capacity to perform certain kinds of actions an *action schema*. For example, a carpenter may share the schema of "making a table". Action schemas, furthermore, have *purposes* and *requirements*, which are other shared schemas. For example, making a table requires the schema "wood" and has the schema "table" as a purpose. On the level of performances, actions can have *results* and *conditions*, which are particular artefacts or actions that may (or may not) actualize these schemas.

Based on this model, we can then define what it means to *succeed* in a very general sense (Figure 3): namely to actualize the purpose of some action schema in terms of the result of performing a corresponding action. Note that actions therefore *can fail* in various ways depending on the circumstances. For example, an apprentice carpenter might use the wrong kind of wood and thus produce a result which is not stable enough to be used as (and thus be actualized as) a table. Or the apprentice might not even be able to make a table in the first place, because the required wood is not available[4]. Note that we can *explain* failure if we lack an artefact that actualizes a requirement. More generally, *explanation*, as suggested by von Wright, reveals the (causal) conditions of processes or states [52]. To *understand* [52] an action, in contrast, means to interpret the purpose of its schema, i.e., to insinuate a motive for action [16][5]. Finally, *learning* proceeds by changing roles in requesting of and giving feedback on the success of actions [16].

2.3.2 Experimental Know-How and Core Concepts in GIS

Experiments, in a pragmatic view, are action schemas whose *starting conditions are controlled* by actions [52, p. 72], while the result is generated by the uncontrolled course of an experimental process, whose outcome is *measured* [16] (Fig. 3). The corresponding notions

[3] This distinction is sometimes called *intentionality* to tell apart spontaneous events from actions.
[4] Janich calls the former kind of success "Erfolg", and the latter one "Gelingen" [16]
[5] This fundamental distinction between explanation and understanding was proposed, e.g., by von Wright in 1971 [52].

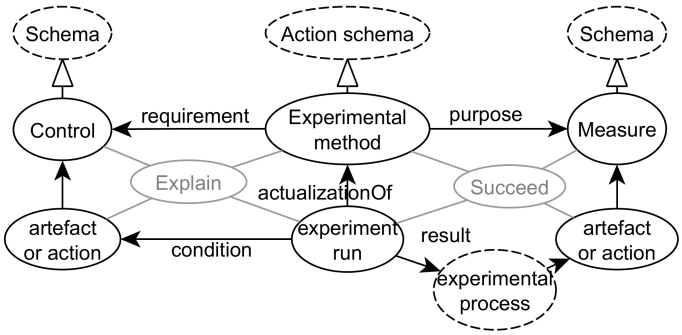


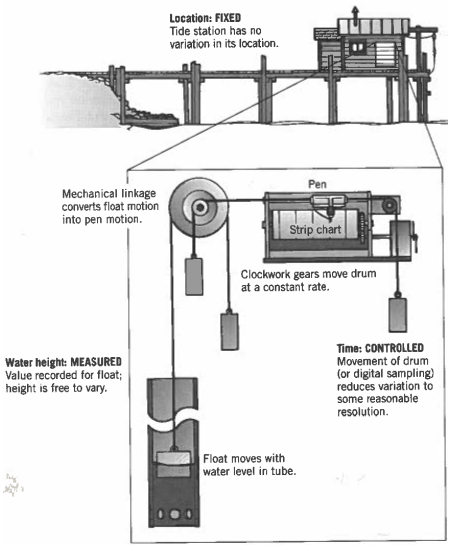
Figure 3 Experimental methods can be understood as kinds of action schemas in Janich's sense [16], which incorporate purposes and requirements, and thus enable success or failure. Experimental methods are action schemas whose requirements are *control* schemas, and whose purposes are *measures* of the result of some (uncontrolled) experimental process which is triggered by the experiment run.

of *control* and *measure*, in this experimental sense, have been suggested by Sinton [41] *to lie at the core of geographic information* [5, ch. 2]. Experimental methods therefore have *control schemas* and *measure schemas*, which are themselves action schemas. An often cited example is a tide gauge (Figure 4a; [5, p. 42]) where the mechanical movement of a drum is used to control time, the location of the gauge controls (and fixes) location, and where the height of water level is measured by a float.

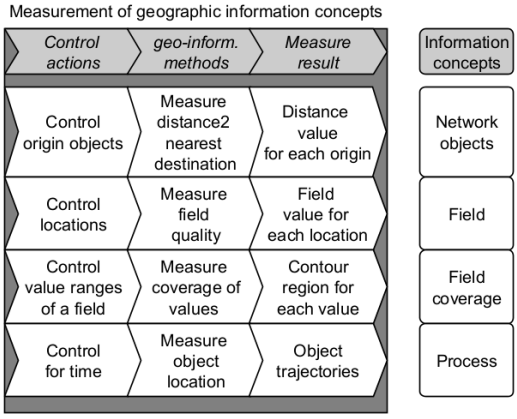
Core concepts of spatial information [19, 49], such as objects, fields, events and networks, as well as their extensive or intensive *quantifications* in terms of amounts or proportions [46], can likewise be understood using an experimental framework (Figure 4b). For example, to assess accessibility, we control for origin objects, and measure travel times, using the concepts of a spatial network and objects [37]. To measure a field like temperature or height, we need to control locations and measure a field quality (Figure 4b), and to measure a contour or a coverage, vice versa, we need to control for field quality ranges, and measure the coverage of those values in the field (e.g. height contours) [38].

2.4 Validity and Purpose of (Spatio-Temporal) Information Methods

So far, we have only talked about the validity of propositions on the levels of data, information and knowledge, and we have introduced a pragmatic account of experimental knowledge. How do information methods fit into this picture? We suggest that *methods in information science*, in general, can be considered action schemas whose *purposes are questions* that need to be answered by the method (they address the information level of the pyramid) (Figure 6). This allows us to make the purposes of methods precise: For example, a *statistical chart* (Figure 5) like a bar chart answers the question "What is the amount of something for each category of a nominal variable"?, whereas a pie chart answers the question "What is the proportion of something for each category of a nominal variable?". In contrast, a histogram answers the question "What is the amount of something for each interval defined on an interval scaled variable?". The validity of a method, in turn, is then defined simply *in terms of whether it can produce a proposition which is a valid answer to this question.* Based on this, we can learn (cf. [53]) that a pie chart is valid only for extensive measures of amounts [46], because measures need to sum up to a total, and that a histogram is an invalid



(a) Geographic information understood as controlling or measuring space or time; example of a tide gauge, see [5].



(b) Examples of spatio-temporal experiments underlying concepts of geographic information, cf. [44].

Figure 4 Examples of spatio-temporal "Sinton" experiments relevant for GIS, together with corresponding information concepts.

method for nominally scaled control variables, because it requires at least interval scaled ones. In general, method applications become unsuccessful in matching their information purpose (2^{nd} criterion above) in case their data requirements are not satisfied.

Similarly, we can understand the validity of GIS methods (and workflows) in terms of the questions they answer [39]. Consider our GIS example about assessing the accessibility of residents for ambulances (Figure 1). First, note that the result of this method is a map that does not directly satisfy this purpose. Instead, it answers a proxy question, namely *what is the distance to the closest ambulance station?* We know this because we use ambulance stations as objects, and because the method used in the workflow measures the network distance to the closest object. We thus know the proxy schema is interpretable (1^{st} criterion). Furthermore, the resulting map should answer the question (2^{nd} criterion), which is only possible if the purpose schema *accessibility of ambulances* can be *inferred* from the proxy schema represented in this map (3^{rd} criterion). This is the case if shortest distance can play the role of accessibility, which indeed corresponds to a practice within the community of geographers [50]. In summary, *validating GIS methods for some purpose* requires interpreting maps as answers to geographic questions [39]. Answers result from a *procedure* (Fig. 6), i.e., from an *answer generation workflow* (cf. [44]). It is for good reasons that such workflows in GIS are often called "models". Yet, models of what precisely?

2.5 Spatio-Temporal Models = Models of Spatio-Temporal Experiments

It is rather uncontroversial that a *model* needs to be *a model* of *something* [9]. In structuralist and realist epistemologies [42, 4], however, it is often claimed that the latter is just "reality". In this view, scientific models allegedly light the way to truth by way of their *resemblance* with reality. This idea has unfortunately become mainstream in the philosophy of science and modeling [11], yet it reflects a rather devastating background ideology [18]. When

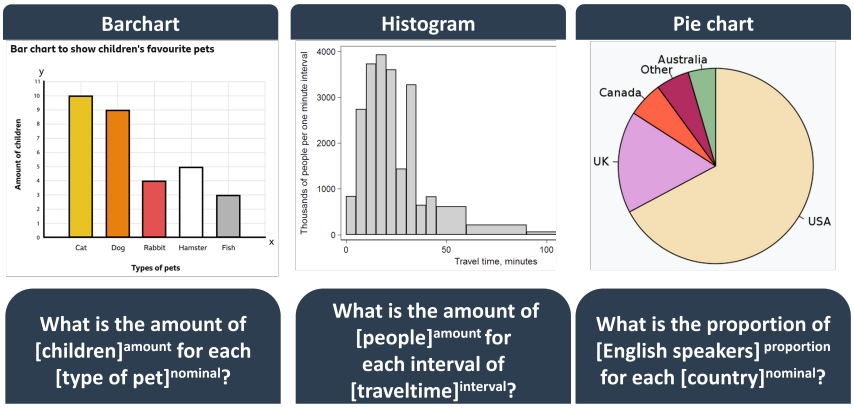


Figure 5 The validity of statistical charts, as information methods, depends on the kinds of questions they can answer about an experiment. In this sense, barcharts are invalid when applied to measures that are not (summable) amounts, histograms are invalid for controls that are not interval scaled, and pie charts become invalid for measures that are not proportions.

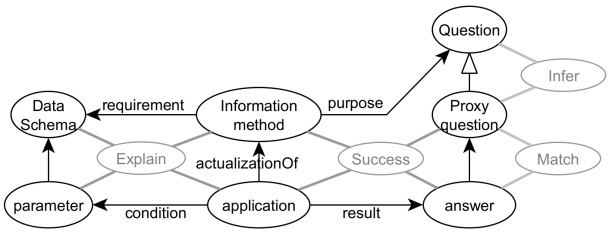


Figure 6 Information methods are action schemas whose purpose is to answer some question. GIS workflows are an example. Often, methods answer only proxy-questions, thus an inference step is needed to determine their validity for a given question (see examples in text).

understanding models in terms of similarity with the real world, serious practical concerns arise [9, 30, 18, 15], in particular when working with simulation models [21]. At the end of the day, reality is not directly accessible, and therefore does not serve as a ground for tests or critique [30, 15]. Furthermore, a simulated reality (a scenario) does not yield any data to test against [21]. Finally, models imply simplification, leaving out aspects that are not considered relevant for the question to be answered by the model [30]. For these reasons, in pragmatic philosophy, truth and reality are not taken as primitive notions, but need to be rooted in validity [15]. And validity, in turn, needs to be rooted in the practice of *justifying* our answers to questions, either by experiment or inference (3^{rd} criterion above).

To account for information models from a pragmatic viewpoint, we therefore suggest to consider them rather as *particular kinds of information methods*, namely ones which *represent experiments*. Experiments are known to be of fundamental importance for the epistemology of simulation models [8, 13]. Since experiments correspond to action schemas (Sect. 3), they can also be *fictive*, in the sense that we can decide to perform or repeat them later. This explains why models are useful, namely as a substitute for experiments which are difficult (e.g. too costly or time-consuming) or undesirable (e.g. the consequences are irreversible) to perform. Models help understand complex systems by playing around with them, which would be impossible in the actual *experiment* they represent. For example, we can change a single driver of the system (control) while holding others constant and measuring the effect on the system state.

Controls and measures have a wider importance for all kinds of experiments in science, and thus also for all models of them. In particular, being able to control actions (both enacting and suppressing starting conditions of experiments [52, p. 72]) is a fundamental prerequisite to determine causality [31], and more generally, for all *explanations* in science [6]. *Natural experiments* control only for starting conditions of any kind of natural process [7]. In *observation* and *measurement experiments* (fundamental for GIS, see [5]), we control for observation time and either location or content (e.g. by choosing objects), while generating values with a sensor or device on a shared scale of measurement. For example, the measurement of a geographic phenomenon such as a temperature field or a the accessibility of a geographic object is originally an outcome of an observation experiment: Of a temperature sensor experiment in the case of temperature, and of a travel time experiment (measuring the time it takes the ambulance to reach you) in the case of accessibility (cf. Figure 7).

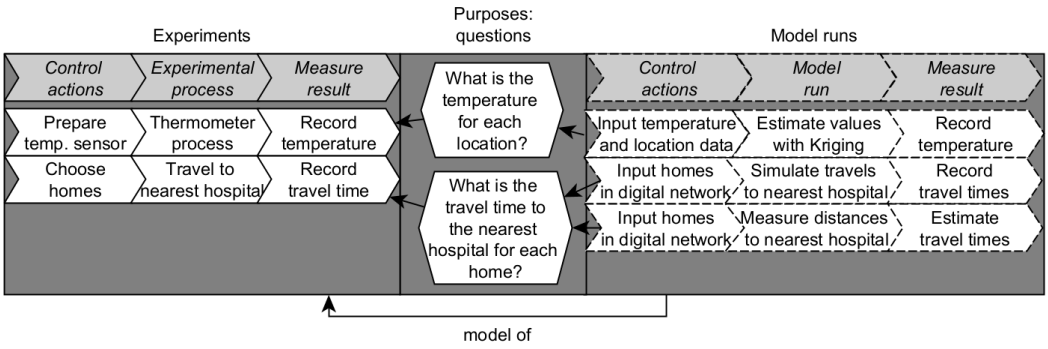


Figure 7 From a pragmatic viewpoint, models are methods that substitute particular experiments which share the same purpose, i.e., they answer the same question. In this way, different kinds of models (from top to bottom: a statistical model, a simulation model, a network model) can be used to model the same experiment.

A *model* of all these experimental methods is just a another method that can be used to *answer the same questions, thus controlling and measuring the same things, but without running the experimental processes* [17]. For example, to save costs, we may run causal models instead of clinical experiments [22]. *Spatio-temporal* models are correspondingly defined as methods that substitute *spatio-temporal experiments*. This latter category includes, in Sinton's sense, all experiments where space and time are among the controlled or measured variables. A *Kriging model of temperature* [20] can play the role of a temperature measurement without running such measurements for all locations in the area of interest (we interpolate a spatial sample). The accessibility workflow in Fig. 1 is, in the most correct sense of the word, a model of the travel times by way of measuring distances on a digital network. Alternatively, we could model this also using a travel time simulation in an agent-based model.

We can differentiate such models according to which kinds of data from experiments are required to build the model: A supervised *statistical model* is one where all variables of the experiment (control and measure) are required *in terms of a sample* from a measurement experiment. For example, in the case of spatial interpolation [20], this sample is obtained from some incomplete measurement of a field over space. A *transformation or inference*

[6] von Wright argued already in 1971 that explanations make use of causality, and that causality can be found, in pragmatic sense, *only based on experimental actions* that can be controlled [52], an argument that has been repeated by modern causal inference theory much later [31].

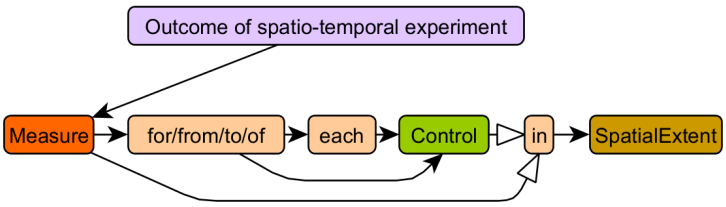


Figure 8 Simplified grammar for a spatial experiment. Black arrows stand for required grammatical choices to build a sentence that describes the experiment, white arrows for optional ones. The "outcome of an experiment" is the starting rule of this grammar.

model [44] (e.g. the accessibility workflow), in contrast, requires no sampling of the measure variables, since those are derived (inferred) from other kinds of measurements. Following the original ideas of Hartmann [13], a *simulation model* (e.g. the travel time agent-based model), instead, is a *model of the experimental process itself.*

3 A Question Grammar Capturing the Purposes of Spatio-Temporal Models

In this section, we develop our method: a grammar centered around the notion of a spatio-temporal experiment, which formalizes the different kinds of purposes of spatio-temporal models in an information pragmatic sense, in terms of different kinds of questions. The grammar can be used to parse questions answered by a given model (section 3.1) and to reason with underlying purposes, which, in turn, points to the possible options for validation (section 3.2).

3.1 Experimental Syntax

The framework of controls and measures not only defines the experiment or a model run, it is also *reflected in the question* that is answered by that experiment or model run (cf. Fig 7). For example, in the question that corresponds to our running example in Figure 1a, we can easily recognize controls and measures (cf. [54]) of the underlying spatio-temporal experiment as follows:

> What is the $^{(Measure)}$[travel time]$_{(interval)}$ to the closest $^{(Control)}$[ambulance station]$_{(object)}$ from each $^{(Control)}$[building]$_{(object)}$ in $^{(SpatialExtent)}$[Rotterdam]$_{(object)}$?

We suggest a grammar that takes *spatio-temporal experiments* into focus as illustrated in Figure 8. Every spatial experiment has some measures, possibly some controls, and some *spatial extent*. This is a way to constrain the experiment within geographic space. In our example, this is given by some spatial object, which is Rotterdam. Furthermore, both the measures and controls correspond to some *concept*, including: *time, interval, location, region, object, quality* (Sect. 2.3.2). This makes it possible to characterize the kinds of experiments modeled based on the kinds of questions answered. In this example, we measure travel time (a time interval) by controlling for origin objects (buildings) within the spatial extent of Rotterdam (another object). Yet, note that we require a *second control* for determining "the closest" ambulance, which needs to be determined relative to each building. How can we account for this?

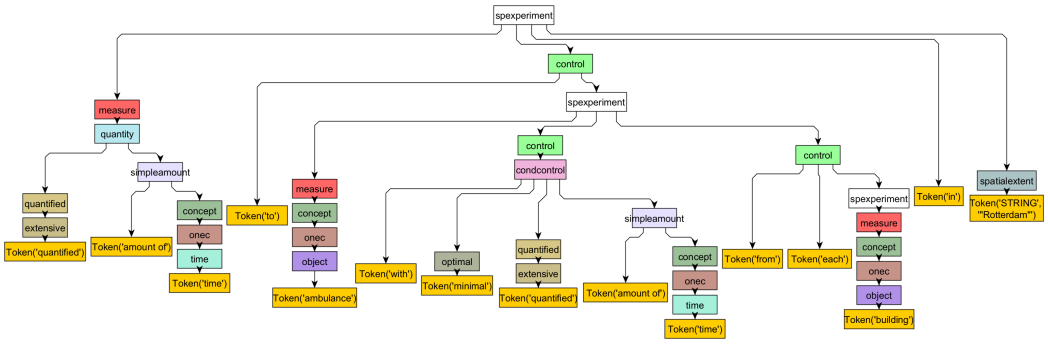


Figure 9 Parse tree of the spatio-temporal experiment denoted by the sentence "quantified amount of time to ambulance station with minimal quantified amount of time from each building in 'Rotterdam'".

A *preliminary formal grammar for spatio-temporal experiments* in EBNF syntax[7], adapted from the work of [54], is given in the Appendix Figure 16a. In this grammar, we used spatio-temporal experiments in a *recursive way* to model arbitrarily complex experimental situations. In our running example, a recursive nesting of experiments as control allows us to capture the meaning of the phrase *"the closest ambulance station"* by reformulating the question in a more explicit manner, in terms of a *nested experiment*. "The closest ambulance station" can then be translated as *"the ambulance station with minimal quantified amount of time"*. This is an *inner experiment*, in which we choose ambulances as measure by controlling the quantified amount of time to reach them, and the entire experiment is then used as a control for the outer experiment, which measures this amount of time using buildings as the control.

> What is the $^{(Measure)}[\text{quantified}]_{(extensive)}$ $[\text{amount}]_{(amount)}$ of $[\text{time}]_{(interval)}$ to the $^{(Control)}[^{(Measure)}[\text{ambulance station}]_{(object)}$ $^{(Control)}\text{with}$ $[\text{minimal}]_{(optimal)}$ $[\text{quantified}]_{(extensive)}$ $[\text{amount}]_{(amount)}$ of $[\text{time}]_{(interval)}]$ from each $^{(Control)}[\text{building}]_{(object)}$ in $^{(SpatialExtent)}[\text{Rotterdam}]_{(object)}$?

The *full parse tree* of this sentence is given in Figure 9[8]. Note that the outer spatial experiment has two nested controls for buildings and their closest ambulances. To simplify the grammar of questions, we allow for abbreviations of such nested experiments using modifiers in the following, like "the closest".

Note that implementing such experiments in terms of measurements, or designing GIS models for transforming geodata sources in terms of a corresponding workflow, requires further procedural knowledge. Automating this task has been called *indirect/geo-analytical question answering* [39]. It was addressed in recent work by parsing questions and querying over automatically synthesized GIS workflow models using *conceptual transformation graphs* [44]. These graphs order the parsed concepts into a tree corresponding to the nesting of experiments, with the outer measure as root (model output) and preceding controls and measures as predecessor nodes, with one or more leafs that correspond to input data [54, 44]. In this way, workflow models for experiments can be automatically constructed for a given question.

[7] `https://en.wikipedia.org/wiki/Extended_Backus-Naur_form`

[8] In the following, we will not display any parse trees due to lack of space. However, all quoted questions are parse-able. The grammar and all parse trees can be found at [36] and `https://github.com/simonscheider/ModelQuestions`.

3.2 Questions as Purposes of Spatio-Temporal Models

So far, our grammar only captures the inner specification of an experiment which might occur in a question. However, the question answered by a spatio-temporal model incorporates also *question words* ("What", "How", etc.), *predicates* ("is"/"are") and *prepositions* ("if"). Furthermore, questions about *modeled experiments* usually are dependent on *modeling conditions*, which form dependent clauses in the question. To capture grammatical differences in modeling purposes, we classify the syntax of questions into five principal types: contemporary inference, prediction, retrodiction, projection and retrojection (Figure 10). Experimental conditions can be captured in terms of further experiments, whose outcomes are constrained by some condition, e.g. by a posited or an obtained result. Regardless of whether *conditioned experiments* are *factual, fictive, or even counter-factual*[9], they are always *causally or quasi-causally* [52] related to the modeled experiments, as reflected in their temporal ordering[10]. This order can be used to distinguish modeling purposes in a principal manner (see *question grammar* in Appendix Figure 16b). Our classification can be used for validation in two ways:

1. To find the right type of model for a certain question (on the *information level*); and
2. To get an overview of the available validations for a certain question-model combination *on the knowledge level*.

In this paper, we focus on the latter. In the following, we explain the five model purposes one by one with parsed example questions. In the next section, we apply the grammar to actually published models and discuss the consequences for validation.

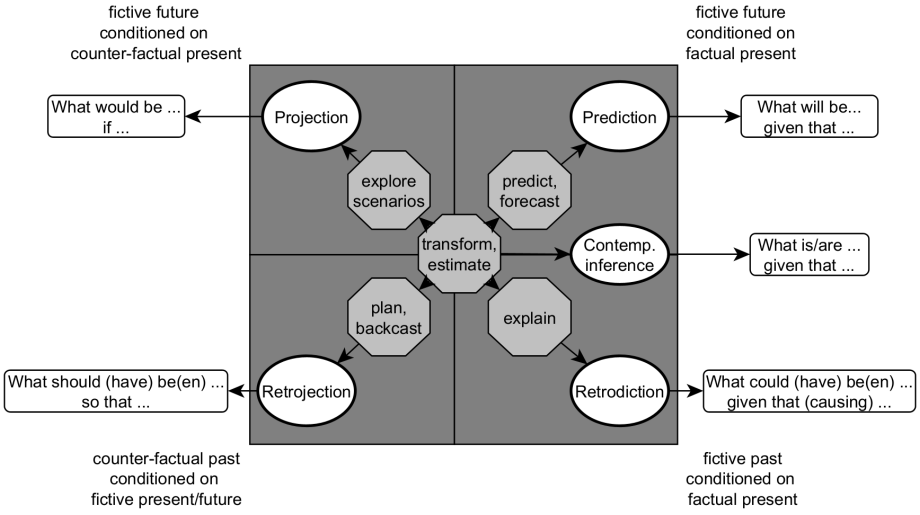


Figure 10 Overview of types of questions considered in this paper. The 4 different planes distinguish types of conditioned experiments in terms of temporal/causal order and degree of fiction.

3.2.1 Contemporary Inference

The first and most simple kind of modeling question asks for a *contemporary inference*. Here, we model a spatio-temporal experiment of the present ("What is/are"), using data of another spatio-temporal experiment of the same time ("given that") (Figure 10, Figure 11).

[9] The latter being fictive experiments with knowingly untrue outcomes.
[10] An experimental cause always needs to precede the outcome of the affected experiment [52].

The first experiment can be fictive, while the latter experiments need to be performed to generate measures to feed the model with some *conditions*. Both experiments have a contemporary reference to time (e.g. "now"). We can further distinguish contemporary

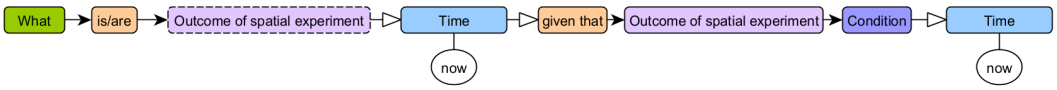


Figure 11 A *contemporary inference model* is used to model a spatial experiment based on performing same (statistical model) or another (transformation model) kind of experiment occurring at the same time. Syntactically, this is expressed by using the present tense "is/are" in the question.

statistical from transformation models, depending on whether the reference experiment is of the same kinds (in terms of action schemas) as the modeled one. For example, a *spatial statistical interpolation model* of temperature may answer the question:

> What is the temperature for each location in Utrecht now, given that the temperature for each sensor location in Utrecht is such and such now?

Note how similar even in the formulation of this question the two experiments are, the fictive one and the factual one.

3.2.2 Prediction

A *prediction model*, in contrast (Figure 10, Figure 12), models a *(fictive) spatio-temporal experiment in the future* ("What will be"), given a factual experiment (usually an observation experiment) at present (see Figure 12). Data from the factual experiment at present is used in order to model the fictive one in the future. Depending on whether these experiments are

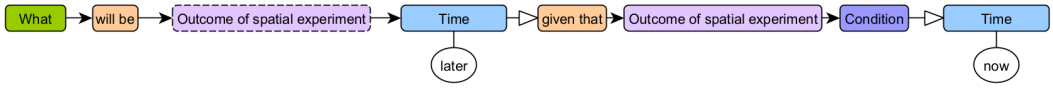


Figure 12 A *prediction model* is used to model a spatial experiment *in the future* based on the same (statistical model) or another (transformation model) kind of experiment occurring at the present time. Syntactically, this is expressed by using the present tense "will be" in the question.

of different or of the same kind, we can draw further distinctions: in the first case, we speak of a *predictive transformation* model. Examples of this can be found in Section 4.2. In the second case, we are concerned with a *predictive statistical* model. An example would be a time series analysis, where we extrapolate a factual time series into the future, as in:

> What will be the temperature in Utrecht tomorrow, given that the temperature in Utrecht is $5°C$ today?

3.2.3 Retrodiction

In a *retrodictive model* (Figure 13), we likewise perform a factual experiment in the present as a condition for a modelled, fictive experiment. However, in contrast to a prediction ((Figure 10, Figure 13), the fictive experiment does not lie in the future, but *in the past* ("What could have been"), thus serving to *explain* ("given that" or "causing") the outcome of the factual experiment performed in the present. For example, we might ask:

> What could have been the event in Utrecht yesterday causing the proportion of water in soil (SWC) in Utrecht being 0.3 today?

Note how the factual experiment is conditioned by a measurement outcome. Retrodictive explanations were formally introduced by von Wright [52, ch. II], who also suggested that such explanations in principle can be done in two different ways, namely based on *causal*

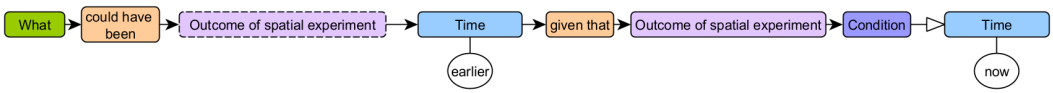


Figure 13 A *retrodiction (= explanation) model* models a spatial experiment *in the past* based on some factual experiment occurring at the present time. The model of the past serves to *explain* the factual, observed experimental outcome of the present. Syntactically, this is expressed by using the present tense "could have been/given that" in the question.

relations (as above), as well as based on inferring the *motives* underlying actions (*teleological explanations*)[11]. In general, retrodictive explanations can be found by *abductive reasoning*, and answer corresponding *why questions* [52, p.94][12]. A retrodiction is therefore far more *remote from experimental validation* than a prediction: If the modeled experiment happens not to have been performed, *it will also never be performed* (as this would require time travel). As a matter of fact, we cannot set up an experiment testing whether a particular decision would lead to the 2nd world war.

3.2.4 Projection

In a *projective* (or *what-if*) model ((Figure 10, Figure 14), in contrast to all previous types, we do not perform a *factual* experiment as an outcome in the present, but rather we *put a condition* on a fictive (*counter-factual*) one. In a projection, this experiment, specified after the phrase "if", serves as a fictive condition for modeling another fictive experiment in the future (specified after "What would be"), whose outcome is the target of the question. The

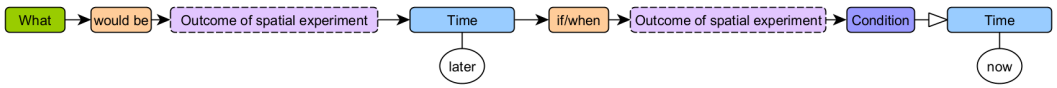


Figure 14 A *projection (= what-if) model* is used to model a spatial experiment *in the future* based on a *model of a counter-factual* experiment occurring at the present time. Syntactically, this is expressed by using the present tense "would be/if" in the question.

outcome of the experiment specified after the *if* clause has a counter-factual outcome at present, meaning it is a *demonstrably false statement*. For example, we might be interested to ask:

> What would be the temperature in Utrecht in 20 years if the amount of CO2 emissions of the world's economy were halved today?

[11] In particular, von Wright [52, ch.III,IV] introduced quasi-causal explanations, which is a mixture of causal and teleological explanations relevant in history and the social sciences. The latter serve to explain events like the 2nd world war by a chain of motives, decisions, and their causal links.

[12] We consider a *why-question* as an alternative formulation of a retrodictive question. Still, we formulate them in terms of *what-questions*, because this form is more explicit and versatile [39, 28]. Note that all question types we introduce here have many alternative formulations.

The importance of asking such questions has been recognized long time ago in climate science, where on a regular basis, projections are used in reports on different climate change scenarios, e.g., as underlying the CMIP6 models of the IPCC [45]. It is no coincidence that the simulation community happens to be among the first to stress the practical importance of *simulation models* for such kinds of questions [51], whereas most (computer) scientists still stick to a generalized, and thus rather superficial, notion of "prediction" [51]. Simulation models answering projective questions are also fundamental for all *planning sciences* [29].

3.2.5 Retrojection

Finally, we suggest *retrojective* models that answer questions about a projection backwards in time: the phrase "What should (have) be(en)" specifies a *fictive experiment in the present (or past)* which is (in contrast to a projection) the target of the modeling (Figure 10, Figure 15). We ask about its outcome under the conditions of a specified outcome of another fictive

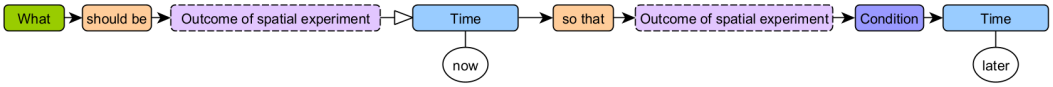


Figure 15 A *retrojection (optimization) model* is used to model a *counter-factual spatial experiment in the present* based on *conditioning* (e.g. constraining, maximizing or minimizing) on the outcome of a model of a fictive experiment in the future. Syntactically, this is expressed by using the expression "should be/so that" in the question.

experiment in the future (phrase "so that"). Thus, retrojective models answer the question how the present needs to be modified in order to reach a certain goal in the future. An example would be:

> What should be the amount of green space in Utrecht today so that the maximum temperature in Utrecht will be below $30°C$ this summer?

The relevance of retrojective questions asking for optimal conditions (in our example above, "so that the maximum temperature ... will be minimal") is apparent in applications of spatial optimization models in planning [29]. However, the more general retrojective question type is also inherent in major recent *research paradigms of planning* framed under the notion of *backcasting* [32], as discussed, e.g., in the context of transformative governance [14, ch.9].

3.3 Evaluation of Grammar

To test our method, in the following, we parse two questions for each modelling purpose: one relatively simple question about the accessibility of ambulances in Rotterdam (cf. Sect. 1) taken from the case study by Geertman et al. [10], and one more complex question derived from a published paper about some model with the corresponding purpose. The first illustrates how a slight change in formulation shifts the purpose and the corresponding options for model validation, whereas the second demonstrates the usefulness of our method to capture the purposes of recently published scientific spatio-temporal models.

4 Illustrations of Modeling Purposes and Assessments of Validity

In this section, we show the results of the evaluation outlined above and discuss the parsed concepts in the tree, as well as what the parsed relation of the pair of involved experiments (cf. Sect. 3.2) tells us about *how the given model could be validated with respect to the question.* Full parse trees are available at [36], in the order of the question occurring in this paper.

4.1 Contemporary Inference

To illustrate contemporary inference models, we focus on our running example again about the accessibility of ambulances. In the study of Geertman et al. [10], the authors used network analysis modeling techniques to answer questions about how to improve the localization of ambulances in Rotterdam. The first goal in this paper targets assessing the *current state* of locations of ambulance stations and how it affects the catchment areas (market areas) for each ambulance station. In our grammar, the underlying purpose, and thus the underlying question, can be formulated as follows:

> What is the closest ambulance station for each building in "Rotterdam" at present given that the location of each ambulance station is such and such now?

According to the question grammar (Figure 11; Appendix Fig. 16b), the question asks about an experiment performed at present that is dependent on the outcome of another experiment performed at present. The grammar (see parse tree) recognized automatically that this first experiment measures a closest ambulance station object controlling for each building object in Rotterdam, given that there is a different kind of experiment determining the locations of the current ambulance station objects. This latter experiment, as a matter of fact, was not modeled, but was obtained by an observation experiment [10], while the first one is the outcome of a spatial network model that took the outcome of the latter experiment as input. This shows that for contemporary inference, only the first experiment is really modeled.

This has important implications for *validation*, which directly follow from the discussion in Sect. 2.2: On the *level of data* (1) we could test the trans-subjectivity of concepts used in the tree. For example, we could check whether our data model is based on the definition of the Regional Ambulance Services (RAVs) in the Netherlands [13]. On the level of *questions* (2), we could check whether the parse trees of questions that our model has been used for in the past match the parse tree of this question in terms of a minimal similarity. On the level of knowledge (3), we can either (3.1) do an experimental validation of the modeled experiment, or (3.2) of the conditioned experiment, or (3.3) we could check the validity of our inference method. The first would require measuring the actual distance from each building to each ambulance station. The second would require repeating measurements of the actual ambulance locations. According to the third option, we would need to test whether shortest paths in our network actually inform us about travel times, which is only an approximation involving assumptions about human travel behavior.

4.2 Prediction

Our standard case published in [10] contains a simple example of a predictive transformation model. The authors used *spatial network analysis (catchment area) models* to predict the expected travel times of ambulances to each building from any existing ambulance station using an up-to-date street network. The corresponding predictive question would be:

> What will be the travel time from the closest ambulance station to each building in "Rotterdam" from now on given that the location of each ambulance station is such and such now?

[13] `https://www.ambulancezorg.nl/en/themes/ambulance-care-in-the-netherlands/regional-ambulance-service/regional-ambulance-service`

, where the term "such and such" refers to the current configuration of ambulance stations as observed now, which corresponds to the outcome of a factual observation experiment different from the one that is modeled.

Predictions using more sophisticated statistical models can e.g. be found in the common scenario of *numerical weather forecast models* [2, 26]:

> What will be the sum of the amount of rain for each location in "Dortmund" tomorrow given that the air pressure and temperature for each location in "Germany" is such and such now?

In this example, the factual contemporary experiment corresponds to air pressure and temperature field measurements over a wider spatial scope, based on which numerical models can predict precipitation amounts over a spatial field [2]. In the parse tree of this question, the spatio-temporal experiments consist in both cases of measures of amounts of stuff controlled by locations, corresponding to the core concept "field". The model is in fact statistical, because it is trained on similar kinds of measurements involving rain gauges. Regarding *validation possibilities*, options (1) and (2) are as above[14]. Regarding the knowledge level (3), at prediction time, the fictive experiment that is modeled in the future cannot be performed, yet (3.1) predictions can always be checked for correctness later by performing *posterior experiments* of the corresponding kind. We can likewise perform repeats of the factual conditioned experiments at present (3.2) and check our inference procedure by testing the underlying transformation or statistical/numerical model separately (3.3).

4.3 Retrodiction

Our standard case of ambulance accessibility contains also an example of a retrodictive model [10]. We may e.g. ask about the reason for the fact that a particular accident in Rotterdam today was in fact, as measured by tracking data, not reached by an ambulance in 14 minutes time, which is the maximal time prescribed by law [10]. We may expect such reasons to lie in the state of the road network at the time, which may have been blocked by the closure of a bridge (an event). Note that in such an explanation, we need to reconstruct possible fictive causes that could have led to the outcome, without necessarily being able to observe them. Yet we might know that the fictive observation experiment of such an event in the past needs to be controlled by a (blocked) pair of road intersections in the street network:

> What could have been the event for each road intersection pair in "Rotterdam" last week causing the travel time of this ambulance station to this road accident in "Rotterdam" to be 30 minutes now?

A more sophisticated model of *historical explanation* [52, ch.], explaining the *current spatial distribution of linguistic groups in the Amazon*, is given by the following example of historical linguistics [27]:

> What could have been the route for each language group in "the Amazon" starting 15.000 years ago causing the location of language groups to be such and such at present?

A possible model that answers this question uses simulated random walks over a phylogenetic tree, making use of Bayesian inference and MCMC sampling to fit the model to observations and to assign a probability to possible evolutionary routes [27]. How could such a model be

[14] Likwise in all succeeding cases.

validated? On the knowledge level (3) the parse tree of this question tells us that there is one factual condition of the present (3.2), which can be tested by repeating the underlying experiment to see whether results are confirmed. In our example, we could e.g. draw new observations about the current locations of language groups. Note, however, how the modeled fictive experiment (3.1) about language origins, since it lies in the past, can never be experimentally validated. Thus the only remaining possibility is to validate our inference methods (3.3) which are used to derive the outcomes of the modeled experiment. In this case, we could e.g. still test whether the underlying simulation rules correspond with what is known about language group behaviors observed in current experiments, see [43].

4.4 Projection

In our standard scenario of ambulance accessibility, we may be interested to know what happens if the road over the Haringsvliet bridge near Rotterdam is closed (compare Sect. 1). As a matter of fact, this bridge is not closed, thus a *counterfactual experiment* can be conditioned by fixing the impedance value of the corresponding pair of road intersections to an arbitrary high value:

> What would be the travel time from the closest ambulance station to each building in "Rotterdam" in the future if the travel time between this intersection pair was infinite minutes from now on?

A more sophisticated example model of a projection was used in [24], where an agent-based model was employed to assess the effects of different policy scenarios for bio-ethanol production (generated from sugar cane) in Brazil. For example, in one such scenario, the authors investigated what effect a consumer tax of 1.23 Real per liter would have on the amount of production of ethanol in the country:

> What would be the sum of the production of ethanol for each producer in "Brazil" in 2030 if the proportional amount of tax for ethanol for each consumer in "Brazil" was equal to 1.23 R/l from now on?

Regarding *validation* on the knowledge level (3), projections are even more remote from experimental validation. Both experiments, the conditioned as well as the modeled one, are fictive, and the first one is even counter-factual. Performing the modeled experiment in the future is *never possible* (3.1) if the conditioned experiment has never been realized, yet counter-factuals may never become real, e.g., in a natural experiment (3.2). Thus the only general way to validate the experiments is to simulate them under realistic conditions. This makes the inference procedure, i.e., in our case the simulation rules (3.3), the only general option for validating a projection. The latter can either be done by calibrating rules in similar experiments [47], or by justifying them based on prior (behavioral) knowledge [1, 43].

4.5 Retrojection

Given that we know, based on prediction or contemporary inference, that Rotterdam's ambulance locations are not sufficient for guaranteeing the 14 minutes threshold, we might be interested in finding out how those locations should be redistributed to achieve this accessibility goal, as expressed by the following parse-able question:

> What should be the location of ambulance stations in "Rotterdam" now such that the travel time to each building from the closest ambulance station will be less than 14 minutes in the future?

Note how in this case, a condition of 14 minutes is placed on a fictive travel time experiment in the future, while the model instead answers a question about some fictive experiment, namely a spatial allocation experiment of ambulance stations in the present. Since this question describes a *constraint satisfaction problem* (= a retrojective question where the condition is a boolean constraint), modeling may consist of a spatial network (catchment area) measure [37] for a limited, manually selected amount of ambulance configurations in GIS, until a satisfactory solution is found.

However, many retrojections require *optimality conditions*, and, thus, *optimization algorithms* to systematically search over the space of possibilities. For example, in [25], the authors presented an optimization model for wind turbine locations (i.e. location allocation) of a wind farm considering the wake effect. The optimization model is based on a genetic algorithm, which searches through possible spatial configurations to answer the following question:

> What should be the location for each windmill of this windfarm now so that the sum of the amount of energy for each windmill of this windfarm will be maximal in the future?

In this study, the authors showed how optimal allocations of turbines can almost double the cost efficiency of energy production of a given windfarm.

Regarding *validation* of a retrojection, the parse tree tells us that there is a fictive conditioned future experiment (3.2), which could be tested later by performing this experiment. For example, we might check whether the windfarm in the future will actually be able to produce the optimal amount of energy. Yet, this will be a validation of the optimization model only in case we are also able to realize the modeled experiment (3.1) (the turbine allocations) at present, to test if it actually leads to optimal outcomes. Thus, experimental validation of this model is, as in the case of a projection, not generally possible. This also leaves us with the validation of the inference procedure (3.3) as the only general option. The latter can be done by justifying both the search algorithm (soundness and completeness), as well as the objective function by inference from further knowledge sources.

5 Conclusion

In this article, we addressed the problem of how to account for the scientific validity of spatio-temporal models more comprehensively, doing justice to the widespread insight that validity considerations should be rooted in the purpose of a model. In the past, it seems that a structuralist background philosophy has upheld progress in these matters, one that has become ubiquitous in modern-day data-driven research [40], and which reduces validations to an exercise in ground truth data comparisons. To overcome this idea, we adopted an approach to validity rooted in pragmatics, and we suggested a corresponding meta-model based on the generalized notion of validity in the pragmatic theory of Janich [16].

In our approach, spatio-temporal experiments consisting of control and measure actions (as introduced in previous work) form the central building blocks of a validity theory. In this theory, validity works on three intertwined levels constituted by corresponding actions: the level of data, constituted of shared interpretations of symbols, the level of information, constituted of question-answering exploiting symbols, and the level of knowledge, constituted of justifications by experiment or inference of concepts occurring in the answers. Correspondingly, questions become the purposes of information models, and spatio-temporal models become models of spatio-temporal experiments that can be validated on each of these

levels. Based on this idea, we proposed a question grammar for spatio-temporal models that distinguishes five different modelling purposes in terms of the types of questions they can answer: contemporary inference, prediction, retrodiction, projection and retrojection.

Using the roles of spatio-temporal experiments in the parse trees of concrete question examples corresponding to these types, we demonstrated that our meta-model can be used to explore the possibilities for validating a spatio-temporal model when applied to a specific purpose. While a given model can serve many different purposes, not all models are fit for the same purpose, and thus can be validated in the same way. This explains why certain modelling communities (like complexity science) have tried to come up with alternative notions of validity that better fit their purpose. However, so far, these endeavours have not led to a useful general concept of validity that subsumes these different purposes, nor to a corresponding theory of validation.

We believe that our meta-model is a start for developing such a theory, having the potential to make the various spatio-temporal modelling purposes explicit. To reach this goal, we plan to further develop and test the grammar, in particular, by combining it with language models to scale it up over multifarious modelling resources. This has numerous potential applications. Questions can e.g. be used to *synthesize workflows* (and thus to automate geographic analysis) for *question-answering* [54, 38, 44], to support *reproducibility* [40], as well as to reason over the *possibilities for validation*. For the latter goal, we would need to link model applications to underlying questions in an extended literature study, which would inform us about the choice of model algorithms for a given question. Specifying experiments in terms of question parse trees then would allow for the assessment of validation possibilities, and for performing validations. For this purpose, our theory needs to be related to data quality and errors models. Given our findings that validation of the inference procedure is the only option for many models, developing specific inference procedures could make explicit the argumentative premise such a model is supposed to support (cf. [47]).

References

1 Craig A Aumann. A methodology for developing simulation models of complex systems. *Ecological Modelling*, 202(3-4):385–396, 2007.

2 Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

3 Hal Caswell. 12: The validation problem. In Bernard C. Patten, editor, *Systems analysis and simulation in ecology*, volume 4, pages 313–325. Academic Press Inc., 1976.

4 Roderick M Chisholm. *Theory of knowledge*, volume 3. Prentice-Hall Englewood Cliffs, NJ, 1989.

5 Nicholas R Chrisman. *Exploring geographic information systems*. Wiley New York, 2002.

6 Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

7 Jared Diamond. Laboratory experiments, field experiments, and natural experiments. *Community ecology*, pages 3–22, 1986.

8 Deborah Dowling. Experimenting on theories. *Science in context*, 12(2):261–273, 1999.

9 Roman Frigg. Scientific representation and the semantic view of theories. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 21(1):49–65, 2006.

10 Stan Geertman, Tom de Jong, Coen Wessels, and Jan Bleeker. The relocation of ambulance facilities in central Rotterdam. *Applied GIS and Spatial Analysis*, pages 215–232, 2003.

11 Ronald N Giere. How models are used to represent reality. *Philosophy of science*, 71(5):742–752, 2004.

12 Lucas Gren. Standards of validity and the validity of standards in behavioral software engineering research: the perspective of psychological test theory. In *Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement*, pages 1–4, 2018.

13 Stephan Hartmann. The world as a process: Simulations in the natural and social sciences. In *Modelling and simulation in the social sciences from the philosophy of science point of view*, pages 77–100. Springer, 1996.

14 Katharina Hölscher and Niki Frantzeskaki. *Transformative climate governance: a capacities perspective to systematise, evaluate and guide climate action*. Springer, 2020.

15 Peter Janich. *Was ist Wahrheit?: eine philosophische Einführung*, volume 2052. CH Beck, 1996.

16 Peter Janich. *Logisch-pragmatische Propädeutik: ein Grundkurs im philosophischen Reflektieren*. Velbrück Wiss., 2001.

17 Peter Janich. Die Naturalisierung der Information. In *Kultur und Methode*, pages 213–255. Suhrkamp, 2006.

18 Peter Janich. *What is information?*, volume 55. University of Minnesota Press, 2018.

19 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276, 2012.

20 Nina Siu-Ngan Lam. Spatial interpolation methods: a review. *The American Cartographer*, 10(2):129–150, 1983.

21 Johannes Lenhard, Günter Küppers, and Terry Shinn. *Simulation: Pragmatic constructions of reality*, volume 25. Springer Science & Business Media, 2007.

22 Matthew McBee. Statistical approaches to causal analysis. *Statistical Approaches to Causal Analysis*, pages 1–160, 2022.

23 Samuel Messick. Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, 14(4):5–8, 1995.

24 Jorge Andrés Moncada, Judith A. Verstegen, John Alexander Posada, Martin Junginger, Zofia Lukszo, André Faaij, and Margot Weijnen. Exploring policy options to spur the expansion of ethanol production and consumption in brazil: An agent-based modeling approach. *Energy Policy*, 123:619–641, 2018.

25 Giovanni Mosetti, Carlo Poloni, and Bruno Diviacco. Optimization of wind turbine positioning in large windfarms by means of a genetic algorithm. *Journal of Wind Engineering and Industrial Aerodynamics*, 51(1):105–116, 1994.

26 Allan H Murphy. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and forecasting*, 8(2):281–293, 1993.

27 Nico Neureiter, Peter Ranacher, Rik van Gijn, Balthasar Bickel, and Robert Weibel. Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? *Royal Society open science*, 8(1):201079, 2021.

28 Enkhbold Nyamsuren, Haiqi Xu, Eric J Top, Simon Scheider, and Niels Steenbergen. Semantic complexity of geographic questions-a comparison in terms of conceptual transformations of answers. *AGILE: GIScience Series*, 4:10, 2023.

29 Stan Openshaw. *Using models in planning: a practical guide*. Retailing and Planning Associates, 1978.

30 Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646, 1994.

31 Judea Pearl. *Causality*. Cambridge University Press, 2009.

32 Jaco Quist. *Backcasting for a sustainable future: the impact after 10 years*. Eburon Uitgeverij BV, 2007.

33 Edward J Rykiel Jr. Testing ecological models: the meaning of validation. *Ecological modelling*, 90(3):229–244, 1996.

34 Gilbert Ryle. *The concept of mind*. Hutchinson, 1949.

35 Simon Scheider. Spatio-temporal modeling questions. Dataset, swhId: swh:1:dir:c63d8dc4808114f58911f1870afaa462b6338d4f (visited on 2024-07-08). URL: https://github.com/simonscheider/ModelQuestions.

36 Simon Scheider. Spatio-temporal modeling questions. Technical report, Utrecht University, 2024. doi:10.5281/zenodo.11066986.

37 Simon Scheider and Tom de Jong. A conceptual model for automating spatial network analysis. *Transactions in GIS*, 26(1):421–458, 2022.

38 Simon Scheider, Rogier Meerlo, Vedran Kasalica, and Anna-Lena Lamprecht. Ontology of core concept data types for answering geo-analytical questions. *Journal of Spatial Information Science*, 20:167–201, 2020.

39 Simon Scheider, Enkhbold Nyamsuren, Han Kruiger, and Haiqi Xu. Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 14(1):1–14, 2021.

40 Simon Scheider and Kai-Florian Richter. Pragmatic GeoAI: Geographic information as externalized practice. *KI-Künstliche Intelligenz*, 37(1):17–31, 2023.

41 David Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*, 1978.

42 Joseph D Sneed. Structuralism and scientific realism. In *Methodology, Epistemology, and Philosophy of Science: Essays in Honour of Wolfgang Stegmüller on the Occasion of His 60th Birthday, June 3rd, 1983*, pages 345–370. Springer, 1983.

43 Tabea S Sonnenschein, G Ardine de Wit, Nicolette R den Braver, Roel CH Vermeulen, and Simon Scheider. Validating and constructing behavioral models for simulation and projection using automated knowledge extraction. *Information Sciences*, 662:120232, 2024.

44 Niels Steenbergen, Eric Top, Enkhbold Nyamsuren, and Simon Scheider. Procedural metadata for geographic information using an algebra of core concept transformations. *Journal of Spatial Information Science*, 27:51–92, 2023.

45 Claudia Tebaldi, Kevin Debeire, Veronika Eyring, Erich Fischer, John Fyfe, Pierre Friedlingstein, Reto Knutti, Jason Lowe, Brian O'Neill, Benjamin Sanderson, et al. Climate model projections from the scenario model intercomparison project (ScenarioMIP) of CMIP6. *Earth System Dynamics Discussions*, 2020:1–50, 2020.

46 Eric Top, Simon Scheider, Haiqi Xu, Enkhbold Nyamsuren, and Niels Steenbergen. The semantics of extensive quantities within geographic information. *Applied Ontology*, 17(3):337–364, 2022. doi:10.3233/AO-220268.

47 Christian Troost, Andrew Reid Bell, Hedwig van Delden, Robert Huber, Tatiana Filatova, Quang Bao Le, Melvin Lippe, Leila Niamir, J Gareth Polhill, Zhanli Sun, and Thomas Berger. How to keep it adequate: A validation protocol for agent-based simulation. *Environmental Modelling & software*, 2023. doi:10.1016/j.envsoft.2022.105559.

48 Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*, 2014.

49 Behzad Vahedi, Werner Kuhn, and Andrea Ballatore. Question-based spatial computing—a case study. In *Geospatial Data in a Changing World: Selected papers of the 19th AGILE Conference on Geographic Information Science*, pages 37–50. Springer, 2016.

50 JR Ritsema Van Eck and Tom de Jong. Accessibility analysis and spatial competition effects in the context of GIS-supported service location planning. *Computers, environment and urban systems*, 23(2):75–89, 1999.

51 Judith Verstegen and Simon Scheider. Why the term prediction is overused. In *Spatial Data Science Symposium 2023 Short Paper Proceedings*. UC Santa Barbara: Center for Spatial Studies, 2023.

52 Georg Henrik Von Wright. *Explanation and understanding*. Cornell University Press, 2004.

53 Leland Wilkinson. *The grammar of graphics*. Springer, 2012.

54 Haiqi Xu, Enkhbold Nyamsuren, Simon Scheider, and Eric Top. A grammar for interpreting geo-analytical questions as concept transformations. *International Journal of Geographical Information Science*, 37(2):276–306, 2023.

A Appendix

⟨spexperiment⟩	⊨	⟨measure⟩(⟨control⟩)* (in ⟨spatialextent⟩)?	(1)
⟨measure⟩	⊨	⟨quantity⟩ | ⟨amount⟩ | ⟨concept⟩	(2)
⟨control⟩	⊨	((for | from | to | of) (each)? ⟨spexperiment⟩) | ⟨condcontrol⟩	(3)
⟨condcontrol⟩	⊨	⟨spr⟩⟨onec⟩ | ⟨compr⟩⟨value⟩ | with ⟨optimal⟩⟨quantified⟩⟨samount⟩	(4)
⟨amount⟩	⊨	⟨samount⟩ | ⟨relamount⟩	(5)
⟨samount⟩	⊨	amount of ⟨concept⟩	(6)
⟨relamount⟩	⊨	amount of ⟨spexperiment⟩	(7)
⟨concept⟩	⊨	⟨onec⟩ | ⟨twoc⟩	(8)
⟨onec⟩	⊨	⟨object⟩ | ⟨event⟩ | ⟨stuff⟩ | ⟨space⟩ | ⟨time⟩	(9)
⟨twoc⟩	⊨	⟨onec⟩ pair	(10)
⟨time⟩	⊨	time	(11)
⟨space⟩	⊨	space | location | height | distance | ...	(12)
⟨spr⟩	⊨	within | touching | away from | west of | ...	(13)
⟨compr⟩	⊨	larger than | less than | equal to | ...	(14)
⟨quantity⟩	⊨	⟨quantified⟩⟨samount⟩ | ⟨aggregated⟩⟨relamount⟩	(15)
⟨quantified⟩	⊨	⟨intensive⟩ | ⟨extensive⟩	(16)
⟨intensive⟩	⊨	proportional | density of | normalized	(17)
⟨optimal⟩	⊨	maximal | minimal	(18)
⟨aggregated⟩	⊨	averaged | ⟨optimal⟩	(19)
⟨extensive⟩	⊨	quantified	(20)
⟨object⟩	⊨	place | building | city | neighborhood | ambulance station | ...	(21)
⟨stuff⟩	⊨	noise | temperature | green | landcover | health | ...	(22)
⟨event⟩	⊨	trip | period | earthquake | ...	(23)
⟨spatialextent⟩	⊨	Rotterdam | ...	(24)
⟨value⟩	⊨	NUMBER UNIT	(25)

(a) Grammar of spatio-temporal experiments. To simplify the grammar, we left away determiners.

⟨question⟩	⊨	(⟨contemprq⟩|⟨prediction⟩|⟨retrodiction⟩|⟨projection⟩|⟨retrojection⟩)("?")?	(26)
⟨condition⟩	⊨	(such and such|⟨optimal⟩|⟨compr⟩⟨value⟩|...)	(27)
⟨factual⟩	⊨	⟨spexperiment⟩(is|are|was|were)⟨condition⟩⟨contemprref⟩	(28)
⟨counterfactual⟩	⊨	⟨spexperiment⟩(was|were)⟨condition⟩⟨contemprref⟩	(29)
⟨projected⟩	⊨	⟨spexperiment⟩(will be)⟨condition⟩⟨futureref⟩	(30)
⟨statistical⟩	⊨	⟨spexperiment⟩⟨contemprref⟩	(31)
⟨transformation⟩	⊨	⟨spexperiment⟩⟨contemprref⟩ given that (the)?⟨factual⟩	(32)
⟨contemprq⟩	⊨	What (is|are) (the)?(⟨statistical⟩|⟨transformation⟩)	(33)
⟨prediction⟩	⊨	What will be (the)?⟨spexperiment⟩⟨futureref⟩given that (the)?⟨factual⟩	(34)
⟨retrodiction⟩	⊨	What could have been (the)?⟨spexperiment⟩⟨pastref⟩given that(the)?⟨factual⟩	(35)
⟨projection⟩	⊨	What would be (the)?⟨spexperiment⟩⟨futureref⟩(if|when)(the)?⟨counterfactual⟩	(36)
⟨retrojection⟩	⊨	What should be (the)?⟨spexperiment⟩⟨contemprref⟩(so|such) that (the)?⟨projected⟩	(37)
⟨contemprref⟩	⊨	now|currently|at present|from now on|at the end of the African humid period|...	(38)
⟨pastref⟩	⊨	earlier|in the past|10.000 years ago|...	(39)
⟨futureref⟩	⊨	in the future|later|in 2030|tomorrow|...	(40)

(b) Question grammar capturing the different purposes of spatio-temporal models.

Figure 16 The grammatical model used in this paper in EBNF syntax (excerpt). The complete grammar and all parse trees available at [36] `https://github.com/simonscheider/ModelQuestions`.