15.03.2025

REPLIKATIONSKRISE

# Hat die Ökologie ein Glaubwürdigkeitsproblem?

Erst erschüttert die Replikationskrise Psychologie und Medizin und jetzt die Ökologie: Fachleute leiten aus exakt denselben Daten völlig unterschiedliche Ergebnisse ab.

von Christian Schwägerl



**Ein Datensatz, viele verschiedene Schlussfolgerungen: Eine solche Replikationskrise zeigt sich nun in der Ökologie.**

Die zwei Fragen klingen harmlos: »Beeinflusst bei jungen Blaumeisen der Wettbewerb mit Geschwistern das Wachstum im Nest?«, lautet die erste. »Wirkt sich die Bodenbedeckung mit Gräsern darauf aus, wie erfolgreich Eukalyptuspflanzen anwachsen?« die zweite. Verbunden mit einem kleinen Datenset oder einer Grafik könnte das so im Biologieabitur oder einer Klausur im Studium drankommen.

In den Fragen steckt aber mehr, viel mehr.

Für eine neu erschienene Studie haben fast 250 Ökologinnen und Ökologen aus aller Welt jeweils identische Daten zu diesen Themen vorgelegt bekommen – verbunden mit der Bitte, diese mit ihren jeweils bevorzugten statistischen Methoden auszuwerten. 146 Teams machten sich unabhängig voneinander an die Arbeit. Eine einfache Sache, sollte man meinen. Muss bei gleichen Daten nicht dasselbe herauskommen? Mitnichten: Die Forscher und Forscherinnen gelangten zu einem breiten Spektrum unterschiedlicher, teils sogar widersprüchlicher Ergebnisse.

»Gleiche Daten, unterschiedliche Analysten: Abweichungen in den Effektgrößen aufgrund analytischer Entscheidungen in Ökologie und Evolutionsbiologie«, heißt die in der Fachzeitschrift »BMC Biology« veröffentlichte Studie. Die Liste der Autorinnen und Autoren erstreckt sich über zwei volle Seiten. Im Kern geht es weder um Blaumeisen noch um Eukalyptus, sondern darum, wie verlässlich die Aussagen wissenschaftlicher Studien sind und wie eindeutig Antworten überhaupt ausfallen können.

## Böse Überraschung beim Ergebnisabgleich

Bei der Blaumeisen-Frage bekamen Forschende, die freiwillig an dem Statistikexperiment teilnahmen, noch unpublizierte Daten über Flügellängen und Körpergewichte von Jungvögeln aus 1100 Nistkästen auf dem Forschungsgelände Wytham Wood der britischen Oxford University übermittelt. Dort hatten Wissenschaftler die Zahl der Nestlinge pro Elternpaar teilweise durch Eingriffe erhöht oder verringert, um Effekte der Geschwisterzahl deutlicher messen zu können. Die Eukalyptus-Daten stammten aus 18 verschiedenen Untersuchungsgebieten im australischen Bundesstaat Victoria, wo Botaniker die Zahl der Jungpflanzen, die Grasbedeckung, Abstände zum nächsten Eukalyptuswald und ähnliche Faktoren erfasst hatten.

Bei der Auswertung der Blaumeisen-Daten kamen fast alle beteiligten Teams zu dem auch intuitiv naheliegenden Ergebnis, dass die Wachstumsraten jedes einzelnen Jungvogels mit der Zahl der Geschwister sinken – schließlich können Eltern nur eine begrenzte Menge Nahrung heranschaffen. Doch die Stärke des Effekts variierte zwischen den einzelnen Analysen teils massiv.

Bei den Eukalyptus-Daten wichen die Ergebnisse viel stärker voneinander ab: Rund die Hälfte folgerte, dass es keinen statistisch nachweisbaren Zusammenhang zwischen der Grasbedeckung und den Überlebenschancen von Sämlingen gibt. Die andere Hälfte der Teams kam zu dem Schluss, dass es Auswirkungen gibt, aber mit teilweise gegenteiligen Aussagen, ob Grasbedeckung die Sämlinge fördert oder unterdrückt.

Schon vor einiger Zeit haben die »Replikationskrisen« in Medizin und Psychologie für große Unruhe gesorgt. Ein wichtiger Aspekt dabei war, dass kritische Forscherinnen und Forscher zeigten, wie viele Studien ein grundlegend wichtiges Merkmal solider Wissenschaft nicht erfüllen: Als man die Studien wiederholte, ließen sich ihre Ergebnisse nicht bestätigen. Bei der Multi-Analysten-Studie liegt der Fall noch extremer: Es geht gar nicht um die Wiederholung ganzer Untersuchungen, sondern um auseinanderklaffende Ergebnisse, die bei der Analyse der exakt selben Daten herauskommen.

## Das Problem ist altbekannt

»Ökologen haben bisher nicht viele Replikationsstudien durchgeführt, um zu quantifizieren, wie groß das Problem bei ihnen ist, aber wir legen sicherlich viele der gleichen Verhaltensweisen an den Tag, die zu dem Problem in der Psychologie und Medizin beigetragen haben«, sagt Rose O'Dea, eine Ökologin von der University of Melbourne.

Einer der Initiatoren und Hauptautoren der Multi-Analysten-Studie ist Tim Parker, ein Biologe vom Whitman College im US-Bundesstaat Washington. Inspiriert sei die Studie aus der Psychologie, sagt er, vor allem von einer Untersuchung aus dem Jahr 2018, bei der zahlreiche Forscher dieselben Daten zu der Frage bekommen hatten, ob Fußballschiedsrichter Spielern mit dunkler Hautfarbe eher Rote Karten geben: »Und wie bei dieser Studie sind auch wir jetzt zu dem Ergebnis gekommen, dass es möglich ist, aus demselben Datenset ein breites Spektrum unterschiedlicher statistischer Aussagen zu generieren«, sagt Parker.

Hat jetzt also auch die Ökologie ein Glaubwürdigkeitsproblem? Die Suche nach Antworten führt tief in den Maschinenraum der Wissenschaft.

Die breite Öffentlichkeit bekommt aus dem Forschungsbetrieb meist nur mit, was ganz am Ende herauskommt – Studien, so genannte Paper, die eine als »peer-review« genannte Begutachtung durchlaufen haben, in einem von tausenden wissenschaftlichen Journals publiziert wurden und über journalistische oder soziale Medien Aufmerksamkeit bekommen.

## »Also wenn einen das besorgt, dann hat man vorher sehr stark die Augen zugemacht«

Florian Hartig, Biologe

Doch lange zuvor treffen Wissenschaftler und Wissenschaftlerinnen weit reichende Entscheidungen: Wie genau lautet die Forschungsfrage? Welche Parameter werden mit welchen Methoden gemessen? Für wie viele Messungen reicht das Budget? Wie genau werden die Daten erhoben? Wie werden die Daten aufbereitet und welche statistischen Methoden kommen zum Einsatz, um sie zu analysieren? Wie interpretiert man die Ergebnisse der statistischen Auswertung? Und zu welchen Antworten auf die Forschungsfrage führt dies schließlich?

Florian Hartig, Professor für Biologe an der Universität Regensburg und dort Leiter der Arbeitsgruppe »Theoretische Ökologie«, hält die Ergebnisse der Studie für »nicht überraschend und auch nicht erschütternd.« Er sieht darin einen eher pädagogischen Versuch, für die Kolleginnen und Kollegen und auch für die Öffentlichkeit noch mal sichtbar zu machen, was aus anderen Feldern der Wissenschaft längst bekannt ist, dass nämlich unterschiedliche Analysetechniken zu unterschiedlichen Ergebnissen führen können. »Also wenn einen das besorgt, dann hat man vorher sehr stark die Augen zugemacht«, sagt Hartig.

### Was man nicht misst, kann richtig viel Ärger machen

Schon wenn Forschende zu wenige Parameter messen, kann sich das rächen und zu hohen Unsicherheiten in den Resultaten führen. Bei Experimenten im Labor sind Wissenschaftlerinnen und Wissenschaftler bemüht, die Bedingungen streng zu kontrollieren, also möglichst wenige und möglichst nur bekannte Faktoren wirken zu lassen. Unter freiem Himmel jedoch, wenn Daten an Brutkästen erhoben werden oder in einer weiten Graslandschaft, ist das unmöglich. In der Natur wirken beinahe unendlich viele, oftmals unbekannte Faktoren zusammen. Neben der Zahl der Geschwister eines Blaumeisenjungen oder dem Grasbewuchs einer Fläche, auf die Eukalyptussamen fallen, reichen zusätzliche Einflüsse vom Wetter über Parasiten über Nährstoffgehalte und Fressfeinde bis zu versteckten Veränderungen im Lebensraum. Zudem ändern sich all diese Faktoren ständig. Freilandforschung hat deshalb eine hohe »Kontextabhängigkeit«, wie Ökologen das nennen.

Wie problematisch es werden kann, zu wenige Aspekte zu messen, zeigt der Insektenschwund. 2017 führte die so genannte Krefeld-Studie zu weltweitem Erschrecken, weil sie darlegte, dass über 25 Jahre hinweg in Deutschland die Biomasse flugfähiger Insekten, die bundesweit über einzelne kürzere Messperioden in Fallen landeten, um 75 Prozent zurückgegangen sei. Umweltschützer und Politiker führten heftige Debatten, welche Änderungen in Landwirtschaft und Naturschutz nötig sind. Zwar liegen viele plausible Änderungen auf der Hand, aber die Studie selbst konnte dazu wenig beitragen. Man habe weder alle Klimafaktoren noch Änderungen in landwirtschaftlichen Praktiken intensiv erfasst, hoben die Autoren der Studie hervor.

Ende 2023 berichtete der Tierökologe Jörg Müller von der Universität Würzburg nach einer eigenen statistischen Auswertung im Journal »Nature« dann, dass sich die Trendzahlen hauptsächlich durch das Wetter erklären ließen. Müllers Studie ist selbst stark umstritten, Kritiker sprechen ihre die Beweiskraft ab und werfen Müller vor, den Effekt der Wetter-Variablen durch methodische Entscheidungen stark zu übertreiben. Die Studie stieß aber eindeutig in eine Lücke der ursprünglichen Datensammlung. Es wäre deshalb wichtig gewesen – und wäre für heutiges Insektenmonitoring wichtig –, möglichst viele Variablen zu erfassen.

Auch dieses Beispiel zeigt: Werden die Daten aus Freilandstudien in die statistische Auswertung gesteckt, ist das eben nicht wie bei einer Kaffeemühle, in die man oben Bohnen reinwirft und unten immer dasselbe Pulver herausbekommt. »Das ist ein beliebtes Missverständnis, wie Wissenschaft funktioniert«, sagt Hartig. In der Realität sind Fachleute vielmehr mit der Aufgabe konfrontiert, Modelle zu erstellen, in denen verschiedene Variablen unterschiedliche Rollen bekommen, zum Beispiel als erklärende Zielvariable, als »Mediatoren«, die Wechselwirkungen verursachen, oder als »Confounder«, die Ergebnisse verzerren können, ohne etwas über Kausalitäten auszusagen. »Und je nachdem, welche Variablen man bei der Auswertung wie berücksichtigt oder nicht, ändern sich auch die Ergebnisse«, sagt Hartig und bezeichnet dieses Vorgehen als »Forscheralltag«.

## Ein Garten voller Fallgruben

Da kann es dann zum Beispiel sein, dass ein Ergebnis statistisch überzeugend aussieht, doch vielleicht nur einen Scheineffekt misst, während ein statistisch schwächer wirkendes Ergebnis den wirklichen kausalen Zusammenhängen eher auf der Spur ist. »Es ist aber klar, mit welcher Analyse man zu einer Publikation in einem Journal kommt«, sagt Hartig – nämlich mit dem Resultat, das möglichst überzeugend und statistisch eindeutig aussieht.

Der Regensburger Biologe verweist auf eine Metapher, die der Statistiker Andrew Gelman von der Columbia University in New York in die wissenschaftliche Methodendiskussion eingebracht hat: Er sprach poetisch von einem »Garten der sich verzweigenden Pfade«, in Bezug zu einem Buch des argentinischen Schriftstellers Jorge Luis Borges. Mit jedem Schritt, den Wissenschaftler und Wissenschaftlerinnen bei der Analyse von Daten gingen – zum Beispiel Variablen ein- oder auszuschließen oder ihre gegenseitigen Einflüsse schwach oder stark zu bewerten –, nehme man eine weitere Abzweigung. »Bei einem Experiment versucht man die möglichen Wege vorzugeben und ihre Zahl kleinzuhalten, aber bei Beobachtungsstudien, deren Bedingungen man nicht kontrollieren kann, geht das nicht«, sagt Hartig, »und bei der Auswertung führen in diesem Garten dann eben nicht alle Wege ans selbe Ziel.« Entscheidend ist es demnach, einen Überblick über die verschiedenen Pfade zu bekommen und ihren Verlauf auszuwerten.

Hat die Studie mit den Blaumeisen- und Eukalyptus-Daten also nur die vielen verschiedenen Pfade im »Garten der sich verzweigenden Pfade« sichtbar gemacht?

> »Richtig problematisch wird es, wenn ein Wissenschaftler ein bestimmtes Ergebnis finden will, fünf verschiedene statistische Modelle ausprobiert und dann das mit dem stärksten Effekt nimmt«
>
> Florian Hartig, Biologe

Ein wirklich hartes Problem entstehe erst dann, sagt Parker, wenn Wissenschaftler es bei einer einzigen Analyse belassen oder – noch schlimmer – aus einem Strauch von Analysen gezielt das Ergebnis herauspicken, das aus ihrer Sicht am besten aussehe und die Aussichten auf eine Publikation erhöhe. »Wenn Wissenschaftler mehrere Analysen durchführen, die dann unterschiedliche Antworten hervorbringen, sie aber nur einen Teil dieser Antworten publizieren, dann ist die Wahrscheinlichkeit groß, dass dadurch eine Verzerrung entsteht, ein

Bias«, sagt Parker. Hartig stimmt zu: »Richtig problematisch wird es, wenn ein Wissenschaftler ein bestimmtes Ergebnis finden will, fünf verschiedene statistische Modelle ausprobiert und dann das mit dem stärksten Effekt nimmt«, sagt er.

»P-Hacking« heißt diese als unseriös verpönte Vorgehensweise in der Statistik, benannt nach dem statistischen p-Wert. Er gibt an, wie wahrscheinlich das gemessene Ergebnis unter der Annahme wäre, dass gar kein Effekt existiert und das Ergebnis nur Zufall ist. Wer P-Hacking betreibt, führt zahlreiche statistische Tests durch, berichtet jedoch nur über jene mit dem Ergebnis, das am überzeugendsten aussieht. Ebenso problematisch ist es, wenn die Datensammlung vorzeitig beendet wird, weil man glaubt, bereits ein aussagekräftiges Ergebnis zu haben, oder wenn man Daten oder Variablen gezielt selektiert, um bei einem bestimmten Resultat anzukommen.

## Falsche Anreize

»Es ist leider ein offenes Geheimnis, dass P-Hacking in der Wissenschaft regelmäßig vorkommt«, sagt Hartig. Dahinter stecke »kein technisches, sondern ein soziales Problem«. Bis heute seien die Anreize in der Wissenschaft nicht dafür ausgelegt, Unsicherheiten bei den Ergebnissen transparent darzustellen. Damit meint Hartig vor allem die Anreize beim wissenschaftlichen Publizieren.

In welchen Journals ihre Ergebnisse erscheinen und wie oft sie zitiert werden, ist für Wissenschaftlerinnen und Wissenschaftler noch immer die härteste Währung. Die Zahl und der »Impact-Faktor« der Publikationen haben einen entscheidenden Einfluss darauf, wer welche Positionen und Förderzusagen bekommt. Statt vielen »Vielleichts« eine klare Botschaft einzureichen, die Herausgeber bei den Journals mit einer knackigen Überschrift versehen können, ist da immer eine große Versuchung.

Was aber könnte und sollte im wissenschaftlichen Prozess angesichts der für viele irritierenden Ergebnisse der Multi-Analysten-Studie besser laufen?

In der Studie selbst heißt es dazu: »Wir sind der Meinung, dass die Ergebnisse einzelner Analysen oder sogar einzelner Metaanalysen in Zukunft mit größerer Skepsis betrachtet werden sollten.« Studien-Koautor Tim Parker zufolge geht es als Lehre genau darum, wie Wissenschaftler in Zukunft offener mit Unterschieden und Unsicherheiten bei ihren statistischen Analysen und deren Ergebnissen umgehen können. Wissenschaftler und Wissenschaftlerinnen könnten künftig in ihren Papern immer einen ganzen Strauch möglicher Ergebnisse präsentieren, statt sich wie häufig auf nur eines zu versteifen.

»Ich würde mir unter uns Ökologinnen und Ökologen eine deutlich stärkere Offenheit dafür wünschen, uns immer wieder selbst zu hinterfragen und dann auch kritisch mit den eigenen Ergebnissen umzugehen – wohl wissend, dass das eventuell einer Publikation in »Science« oder »Nature« im Weg stehen kann«, sagt Ingolf Kühn, Professor für Makroökologie an der Martin-Luther-Universität Halle-Wittenberg und Leiter des Departments Biozönoseforschung am Helmholtz-Zentrum für Umweltforschung (UFZ).

Er hält ein ganzes Bündel von Maßnahmen für erforderlich, um die Qualität wissenschaftlicher Ergebnisse zu erhöhen: Es brauche zum einen einheitlichere Standards, wie Daten erfasst werden, damit sie später besser in gemeinsame Analysen oder Metaanalysen eingehen können. Kühn sieht hier die bisherige Praxis in den USA als Vorbild an, wo es bei vielen Vegetationsstudien landesweite Vorgaben gibt, was und wie es erfasst wird. »In Europa dagegen hat jedes Land und oftmals jede Forschergruppe ihre eigenen Methoden, und anschließend ist es ein riesiger Aufwand oder sogar unmöglich, die Daten vergleichbar zu machen«, sagt Kühn. Sinnvoll sei es auch, Forschungsprojekte vorab registrieren zu müssen,

dabei alle Ziele und Methoden zu hinterlegen und später so genannte Registered Reports zu publizieren: »Die Registrierung schreckt nachweislich davon ab, später Ergebnisse schönzurechnen«, sagt er.

## Viele Ergebnisse statt einem einzigen

Als weitere große Schritte hin zu einer besseren Forschungspraxis hält Kühn die Bereitschaft für wichtig, eine viel größere Bandbreite statistischer Ergebnisse zu präsentieren und insgesamt größtmögliche Offenheit zu praktizieren. »Wir sind immer noch nicht gut darin, neben den Daten auch die Codes offenzulegen, mit denen wir arbeiten«, sagt Kühn. Er habe 14 Jahre als Chefredakteur eines wissenschaftlichen Journals gearbeitet und dabei, wie er sagt, »noch kein einziges Methodenkapitel gesehen, bei dem man zu 100 Prozent wusste, was gemacht worden ist, wo welche Entscheidungen aus welchen Gründen getroffen worden sind und wie ich das als Außenstehender wieder in Code übersetzen kann«.

> »Natürlich möchte jeder ein tolles Paper in einer tollen Zeitschrift haben, aber zu oft ist der Wunsch, da etwas unterzubringen, größer als der strenge wissenschaftliche Anspruch – das muss sich ändern«
>
> Ingolf Kühn, Makroökologe

Kühn erkennt in der Blaumeisen-Eukalyptus-Studie eine Mahnung, »von einem Wissenschaftsverständnis wegzukommen, wo man nur Professor wird, wenn man viele »Nature«- und »Science«-Paper vorweisen kann«. Belohnt werden solle wissenschaftliche Rigorosität – wozu Kühn es auch zählt, Unsicherheiten und alternative Analyseergebnisse zu benennen: »Natürlich möchte jeder ein tolles Paper in einer tollen Zeitschrift haben, aber zu oft ist der Wunsch, da etwas unterzubringen, größer als der strenge wissenschaftliche Anspruch – das muss sich ändern.«

Florian Hartig sieht auch die wissenschaftlichen Journale in der Pflicht: »Eine praktische Möglichkeit wäre, dass sie Daten und Analyseskripte zu jeder Publikation verfügbar machen«, sagt er. Besser sollten auch die Möglichkeiten von Forscherkolleginnen und -kollegen werden, alternative Analysen von Daten direkt neben der Originalstudie zitierbar zu veröffentlichen. »Im Moment schreibt man ja oft eine Response mit einer alternativen Auswertung, die wird dann aber nie veröffentlicht, weil das Journal das ablehnt«, sagt er. Auch die Vorab-Registrierung von Forschungsvorhaben sieht Hartig positiv: »Es sollte nicht zu rigide werden und die Forschung nicht einschränken, aber es ist belegt, dass eine Vorab-Registrierung die wissenschaftliche Praxis verbessert und P-Hacking reduziert.«

Mehr Transparenz und Offenheit in der Forschung ist seit 2020 das Ziel einer eigens gegründeten Organisation, der »Society for Open, Reliable and Transparent Ecology and Evolutionary Biology«, aus deren Umfeld einige der Hauptautoren der Multi-Analysten-Studie kommen. »Sortee« hat bereits Mitglieder aus aller Welt, die in Ökologie und Evolutionsbiologie die Diskussion über Methoden und Standards voranbringen wollen. Die Wissenschaft müsse die Anreize für Forscherinnen und Forscher reduzieren, sich zum Publizieren »die überzeugendste Geschichte herauszupicken«, sagt Ökologin Rose O'Dea, die die Organisation mit aufgebaut hat. Wie Kühn und Hartig sieht sie es als ersten Schritt an, dass »Autoren ihre Daten und ihren Code für alle zugänglich machen und nicht nur ihre Ergebnisse«. Sie nennt so genannte verblindete Datenanalysen als weitere mögliche

Verbesserung. Dabei nutzen die Forschenden ihre Variablen in der Auswertung ohne Label – und können nicht solche Ergebnisse herauspicken, bei denen ihre Zielvariable besonders wichtig erscheint. Das Risiko dabei ist allerdings, dass man am Ende nur Scheinkorrelationen bekommt, weil das Modell eben auch gegenüber bekannten Kausalitäten in der Natur blind ist.

Der aktuelle Präsident von »Sortee«, Ed Ivimey-Cook von der University of Glasgow, hält die Multi-Analysten-Studie für einen Weckruf: »Wir müssen weg davon, die Menge der Publikationen eines Forschers in den Fokus zu stellen, und viel stärker auf die Qualität schauen«, sagt er. Diese zeige sich nicht in schlagkräftigen Überschriften von Studien, sondern darin, »wie offen und selbstkritisch Forschende mit Unsicherheiten in ihren Ergebnissen umgehen«.

**Christian Schwägerl**

Der Autor ist Journalist, Buchautor und Mitgründer von »RiffReporter«. Von ihm stammen die Bücher »Menschenzeit« über das Anthropozän, »11 drohende Kriege« über globale Konfliktrisiken und »Die analoge Revolution« über die Zukunft digitaler Technologien.