RETHINK RESEARCH

Synthetic Populations in Market Research

INTRODUCTION

SYNTHETIC POPULATIONS: WHY THEY MATTER NOW

Organizations are under pressure to cut costs and improve efficiency. Decisions must be made faster, and that requires a solid foundation of data.

However, traditional market research reaches its limits when it comes to highly fragmented markets, very small or niche target groups, or audiences confined to specific regions. Time and budget constraints often undermine early hypothesis testing, while data protection and transparency remain non-negotiable.

This is where synthetic populations can serve as a valuable complement to traditional market research.

Eike Hartmann
Senior Director
Data Insights

E-Mail ↗

Claudia Cramer
Senior Director
Market Research
Insights

E-Mail ↗

statista+

Statista+ enhances Statista's global data platform with a comprehensive suite of services: custom market research and analytics, strategy consulting and business-building, as well as design and marketing solutions, tailored to the specific needs of clients across a broad range of industries worldwide. With a team of over 200 experts from diverse disciplines, Statista+ empowers organizations to unlock their full potential through tailored, data-driven solutions.

SYNTHETIC DATA

BEYOND THE HYPE: AN OVERVIEW

From synthetic panels and persona bots to synthetic populations, synthetic data is currently in the spotlight of market research and is being discussed intensively. This paper provides guidance and puts the focus on synthetic populations and how they are reshaping the practice of market research.

In essence, synthetic populations are not a new concept. In academic research and experimental statistics, they have long been used for simulations and modeling. What is new is their growing relevance for market research.

To set the stage, we distinguish synthetic populations from surrogate data and synthetic respondents:

Surrogate data

Definition: A very large, detailed original dataset (together with its statistical patterns) is used to train a model that generates a fully synthetic dataset closely resembling the original data.

Use: Analysis of highly regulated data, for example, in the health sector. One example is synthetic datasets composed of realistic but nonreal records that mirror enrolment information and healthcare services for Medicare beneficiaries.

Data quality	●●●●
Data depth	●●●●
Data flexibility	○○○○
Effort to generate	●○○○

Synthetic populations

Definition: Realistic, privacy compliant representations of real populations (“digital twins”), where attributes such as age, gender, income, household composition or place of residence follow the statistical patterns of the actual population (e.g. marginal distributions).

Use: Representative analyses, micro segmentation, and scenario modeling at population level.

Data quality	●●●○
Data depth	●●●●
Data flexibility	●●●○
Effort to generate	●●●●

Synthetic respondents

Definition: Simulated primary market research; AI-supported simulated responses or virtual participants for hypothetical research questions.

Use: Early hypothesis testing, ideation, and concept exploration.

Data quality	●○○○
Data depth	●●○○
Data flexibility	●●●●
Effort to generate	●○○○

SYNTHETIC DATA

The middle way: High realism with necessary flexibility

Synthetic populations combine high realism and data protection with the flexibility required for audience specific analyses and scenario calculations, for example price changes. They are not generated by generative AI, but are created by human experts using multiple data sources and statistical modeling.

The result is a robust, representative foundation for segmentation and simulation, for example, of consumer behavior. For many use cases in audience analysis, this middle way proves particularly resilient.

The two extremes: Unusable data assets and flexibility with uncertain quality

Surrogate data is the historically oldest approach. It offers high quality and depth with relatively low effort to produce, but is usually difficult to access and extremely inflexible in practice.

Synthetic respondents – including synthetic panels and persona bots – are fully AI driven, highly flexible, and inexpensive to generate. However, without proper validation, the data quality remains uncertain.

SYNTHETIC POPULATIONS

Old concept, new opportunities in market research: Synthetic populations are now specifically optimized for audience analysis.

High realism through a rich data foundation: Official statistics, surveys, observational data, transaction data, etc.

Flexible output: A representative, finely segmental picture of the population, without any data created by generative AI.

SYNTHETIC POPULATIONS

Fast tests, more depth, population wide answers

Where representativeness and granularity matter, synthetic populations are setting new standards – and are already used successfully in practice.

MOBILITY PLANNING IN SWITZERLAND

Since 2014, the Swiss Federal Office for Spatial Development (ARE) has been creating synthetic populations as a basis for mobility planning.

Traffic flows and new mobility offerings can thus be simulated in a realistic way, including for rare target groups.

DIGITAL TWINS IN THE UNITED KINGDOM

Universities and consortia (including Leeds, Glasgow, CDRC, SIPHER) use synthetic populations to model “digital twins” for research on social and health topics.

This allows them to simulate the effects of policy measures and broader societal developments.

DIGITAL REPRESENTATIONS OF REAL SOCIETIES

Synthetic populations are artificially generated datasets that reproduce a real population (e.g., of a country or specific region). To do this, profiles of individuals or households with attributes such as age, gender, income, household size and place of residence are combined in such a way that they match official statistics and survey data.

In this way, the synthetic population reflects census or survey data. The focus is on patterns, not on real people. No personal data is required or used. The result is a robust basis for analyses at population level, without any data protection risks.

A MATTER OF QUALITY

HOW "REAL" IS A SYNTHETIC POPULATION?

Synthetic populations are modeled representations of reality. They are created by transferring statistical patterns – from sources such as official statistics, panels, and surveys – into probability distributions that are applied to virtual individuals. In doing so, the synthetic individuals replicate the structural characteristics and relationships found in the real population.

Whether the resulting insights hold up depends on a few, but decisive factors: the quality and breadth of the underlying data, how biases are handled, the type of research question, and the recognition that human behavior can never be fully reproduced.

In practice, this means: synthetic populations are a powerful accelerator and a quality boost – not a replacement for traditional market research. They help to narrow down ideas quickly, identify priorities, and reserve time and budget for questions that truly require real human feedback.

TRANSPARENT, HIGH-QUALITY DATA FOUNDATION

The strength of any results stands and falls with the underlying data. If sources are broad, up to date, and representative, the population's picture will come close to reality; if they are narrow or patchy, reliability declines. To be able to judge this, transparency about data provenance is crucial. In short: "Garbage in, garbage out," remains the basic rule.

BEHAVIOR IS NOT 1:1 REPLICABLE

Irrational or culturally shaped behavior does not strictly follow statistical rules. Models can approximate reality, but they do not replace real observation. Results should therefore always be interpreted in context.

RECOGNIZING AND LIMITING BIASES

Data carries its own history. If the input data contains biases (e,g., under-representation of certain groups), these can be perpetuated. Bias checks, documentation, and, where possible, corrections are therefore an essential part of the process.

LIMITS FOR EXPERIENCE- AND EMOTION-DRIVEN QUESTIONS

Not every research question is equally well suited. Topics that are strongly influenced by emotions, individual experiences, or situational nuances can only be modeled to a limited extent. In such cases, traditional research remains the method of choice.

NEW POSSIBILITIES

WORKING WITH SYNTHETIC POPULATIONS

From rare segments to regional drill downs, synthetic populations provide immediate, granular, and valid data without personal information.

They offer reliable magnitudes, profiles, and regional differences that traditional approaches can only capture slowly, or in some cases not at all.

Making rare target groups visible

? Panels reach their limits when trying to determine the size, profile, and distribution of rare groups, such as people with physical impairments or a migration background.

✓ Synthetic populations make it possible to generate robust indications for product design, targeting, and accessibility.

Analyzing fragmented markets and niche potential

? Products now address a wide range of micro-segments. Region, life stage, and context make a difference. “Average values” are no longer enough.

✓ Synthetic populations provide a population-wide, finely resolved view down to the very smallest segments and regions, and show potential for special-interest markets.

Audience validation before large studies

? Early hypothesis testing before expensive segmentation work often fails due to time and budget constraints.

✓ Synthetic populations make it possible to run “slice and dice” analyses at a regional level, estimate reach, and sharpen sample design. Large studies can then start more focused and efficiently.

Regional drill-downs

? Which local factors such as infrastructure, demographics or income influence growth? Where is interest and purchase intent highest?

✓ With synthetic populations, market questions can be answered at a very granular level; answers become comparable, nationwide, and locally differentiated.

STATISTA SYNTHIEPOP

STATISTA SYNTHIEPOP: PRACTICE AND OUTLOOK

From principle to platform: academic synthetic populations are usually based solely on official data (e.g., census data) and pursue a clearly defined public purpose, such as mobility research, public policy, health optimization or pandemic simulation.

Statista is developing a flexible synthetic population for Germany for market research purposes: ***Statista SynthiePop.*** Statista SynthiePop also uses official statistics as a foundation, but enriches them with proprietary Statista data, including market research data and modeled variables.

The result is significantly higher flexibility and a much better fit for market research, especially for measuring and profiling target groups. In addition, the system is flexibly extendable, meaning that new attributes can be added to virtual individuals where this is meaningful and methodologically sound.

What Statista SynthiePop offers

Statista SynthiePop is a synthetic population that represents the entire population of Germany. Because all data is available at the level of synthetic individuals, all dimensions can be cross-tabulated at any level.

The attributes are derived from several official sources (such as government statistics), from proprietary Statista data, and from modeled features where appropriate and methodologically justifiable.

Fully compliant with data protection requirements: no inferences about real persons and a 100% anonymized synthetic basis.

What Statista SynthiePop can be used for

- ***Drill downs*** to very fine regional levels and slice and dice analysis across freely selectable attributes.
- ***Identification and quantification*** of niche and hard to reach segments at national level.
- ***Preparing and validating large segmentation studies:*** estimating reach, sharpening sample designs, and prioritizing hypotheses.

A STRONG COMPLEMENT

From quick pre-analysis to in-depth audience profiling, Statista SynthiePop delivers precise reach estimates, profiles and regional insights without immediately having to launch a new, expensive primary study.

Case Example

The research question

How does the "Deutschland Ticket" (Germany's nation-wide public transport ticket) affect commuting behavior as well as perceived and actual commuting costs?

Existing studies did not provide detailed insights into specific commuter segments, cost drivers, and changes in mobility behavior over time, especially regarding different distances, regions, and combinations of modes of transport.

The solution

Combining data: A dedicated survey of 5,000 commuters, relevant mobility studies, data on commuting behavior from the German Federal Employment Agency, and Statista Consumer Insights.

Mapping to the population: All information for more than 40 million commuters at individual level was projected onto the synthetic population of Germany.

Analyzing the synthetic population: Comparison of car versus public transport usage by region, commuting distance and profile, including cost effects and behavioral changes.

The result

Behavioral change: Shifts in commuting behavior, such as changes in mode of transport, become visible at regional level.

Financial impact: It becomes clear which commuter groups and regions benefit, and to what extent.

Decision relevance: The results provide a robust view at population level, differentiated by region and target group, as a basis for service planning, communication, and infrastructure decisions.

STATISTA SYNTHIEPOP

CASE EXAMPLE: THE "DEUTSCHLAND TICKET"

Analysis of commuting behavior and cost effects for the German Centre for Rail Traffic Research at the Federal Railway Authority.

FUTURE OUTLOOK

WHERE DO WE GO FROM HERE?

A synthetic population that covers the entire population can, in combination with real market research data, form a strong basis for target audience analyses or simulations. In this way, it complements traditional market research and can partially replace it.

Synthetic populations make research not only faster but also more precise and interactive. Tables become dialogues; snapshots become scenarios.

What does this mean in practice? Four application levers with Statista SynthiePop:

EXTEND

Integrate your own customer or survey data into Statista SynthiePop and analyze results in full granularity.

INTERACT

Bring target audiences to life: use Statista SynthiePop as a basis for qualitative primary research and interview synthetic representatives (persona bots) to assess reactions to new concepts.

REPLICATE

Use Statista SynthiePop for quantitative primary research and simulate quantitative surveys with a better representation of the total population, for example to complement hard-to-reach target groups.

SIMULATE

Run what-if scenarios in minutes, model population scenarios, and analyze changes such as shifts in purchasing behavior or mobility patterns.

DO YOU HAVE QUESTIONS?

If you would like to learn more about how synthetic populations can be integrated into your projects, or if you are interested in testing the Statista SynthiePop beta version, please feel free to contact us for a non-binding conversation.

contact@statistaplus.com

Statista+ | Statista SynthiePop Team
statistaplus.com