5 Datenaufbereitung und Berechnungsmethode

5.1 Validierung, Imputation und Plausibilisierung

Die Validierung und Plausibilisierung der Daten sind sehr wichtige Schritte für die Qualitätssicherung sowie zur Beseitigung falscher oder unwahrscheinlicher Transaktionen. Einige Lücken (fehlende Werte) können dank Imputation geschlossen werden. So wird verhindert, dass Transaktionen, die ansonsten von guter Datenqualität sind, ausgeschlossen werden müssen.

Eine erste Validierung erfolgt mit der IT-Anwendung (BFS-IT-Modul), die für die Anreicherung und Anonymisierung der Daten verwendet wird (vgl. Kapitel 3). Die Anwendung validiert das Format und die Werte der Variablen. Auch die Qualität des Adressabgleichs wird kontrolliert. Anhand einer Logdatei werden die Datenlieferanten gegebenenfalls über Fehler informiert und haben so die Möglichkeit, die Daten vor der Übermittlung an das BFS zu korrigieren.

Nach der Zustellung der Daten an das BFS werden sie vom BFS-Informatiksystem erneut validiert und analysiert. Dabei werden die Anzahl gemeldeter Transaktionen im Vergleich zu den vorangehenden Quartalen, die Qualität der Anreicherung sowie allfällige Dubletten und Extremwerte überprüft. In dieser Phase ist eine enge Zusammenarbeit mit den Datenlieferanten erforderlich, damit die Qualität der Beobachtungen gewährleistet werden kann.

Bei der Validierung der Daten kann sich herausstellen, dass im übermittelten Datensatz zu einer Transaktion eine Variable fehlt, beispielsweise die Anzahl Zimmer oder Badezimmer. Damit die Transaktion berücksichtigt wird, müssen jedoch sämtliche vorab definierten und erforderlichen Variablen vorhanden sein. Daher werden fehlende Variablen wenn möglich imputiert (geschätzt). Die Imputationsregeln wurden wie folgt festgelegt:

- Kaufpreis: keine Imputation möglich, die Transaktion wird entfernt
- Transaktionsdatum: Imputation des aktuellen Erhebungsquartals
- Objekttyp: Imputation anhand der Variablen «Einfamilienhaustyp», «Eigentumswohnungstyp», «Kubatur» und «Nettowohnfläche»
- Baujahr: Imputation der Durchschnittswerte (arithmetisch) der Zelle, zu der die Transaktion gehört
- Grundstücksfläche (bei Einfamilienhäusern): Imputation eines Werts anhand der Kubatur und des Einfamilienhaustyps
- Kubatur (bei Einfamilienhäusern): Imputation eines Werts anhand der Grundstücksfläche sowie der Anzahl Zimmer und Badezimmer
- Norm für die Kubaturmessung (bei Einfamilienhäusern): Imputation der Durchschnittswerte (arithmetisch) der Zelle, zu der die Transaktion gehört
- Nettowohnfläche (bei Eigentumswohnungen): Imputation eines Werts anhand der Anzahl Zimmer und Badezimmer
- Anzahl Zimmer: Imputation eines Werts anhand der Kubatur (bei Einfamilienhäusern), der Nettowohnfläche (bei Eigentumswohnungen) und der Anzahl Badezimmer
- Anzahl Badezimmer: Imputation eines Werts anhand der Kubatur (bei Einfamilienhäusern), der Nettowohnfläche (bei Eigentumswohnungen) und der Anzahl Zimmer
- Bauqualität: Imputation eines Werts anhand des Gebäudezustands, des Baujahrs und der durchschnittlichen Fläche pro Zimmer (bei Eigentumswohnungen)
- Gebäudezustand: Imputation eines Werts anhand der Bauqualität und des Baujahrs

Sobald die fehlenden Variablen imputiert sind, werden nicht plausible Wertkombinationen und Extremwerte ermittelt und entfernt. Ein Extremwert ist falsch, unwahrscheinlich oder überschreitet bestimmte festgelegte Grenzen. Er kann die Ergebnisse wesentlich verfälschen, weshalb dieser Verarbeitungsschritt von zentraler Bedeutung ist.

Um die Plausibilität der Variablenwerte einer Transaktion innerhalb des hedonischen Modells zu prüfen, wird die sogenannte «Cook's Distance»-Methode beigezogen.[21] Mit dieser Methode werden Ausreisser und nicht plausible Kombinationen von Werten, die für sich genommen zulässig sind, ermittelt und die entsprechenden Transaktionen entfernt. Beispielsweise würde eine Transaktion für eine 7-Zimmer-Wohnung mit 25 m^2 Nettowohnfläche als nicht plausibel eingestuft, da die Kombination dieser beiden Variablen (einzeln plausibel) sehr unwahrscheinlich ist.

[21] Die Cook's Distance wird in der Statistik eingesetzt, um den Einfluss einer multivariaten Beobachtung im Rahmen einer Kleinste-Quadrate-Regression zu schätzen. Das Konzept wurde erstmals vom amerikanischen Statistiker R. Dennis Cook vorgestellt (Detection of Influential Observation in Linear Regression, Technometrics, Band 19, Nr. 1, Februar 1977). Die Cook's Distance misst den Effekt der Auslassung einer Beobachtung im hedonischen Modell. Ausreisser mit hohen Residuen und/oder einer starken Hebelwirkung auf das Modell können das Ergebnis verfälschen und die Genauigkeit der Regression beeinträchtigen. Transaktionen mit einer grossen Cook's Distance gelten als nicht plausibel und werden bei der Berechnung des Schweizerischen Wohnimmobilienpreisindex ausgeschlossen.

Entfernt werden auch Transaktionen mit Extremwerten, d.h.:

- Transaktionsdatum ausserhalb der Erhebungsperiode (Quartal)
- Kaufpreis < 100 000 Fr. (bei Einfamilienhäusern) bzw. < 75 000 Fr. (bei Eigentumswohnungen) oder > 10 000 000 Fr.
- Grundstückfläche < 50 m^2 oder > 5000 m^2 (bei Einfamilienhäusern)
- Kubatur < 200 m^3 oder > 3000 m^3 (bei Einfamilienhäusern)
- Nettowohnfläche < 20 m^2 oder > 300 m^2 (bei Eigentumswohnungen)
- Anzahl Zimmer < 1 oder > 15 (bei Einfamilienhäusern) bzw. > 12 (bei Eigentumswohnungen)
- Anzahl Badezimmer < 1 oder > 8 (bei Einfamilienhäusern) bzw. > 6 (bei Eigentumswohnungen)

Diese Extremwerte wurden auf Basis der Transaktionen aus drei Jahren (2017 bis 2019) bestimmt. Zusammengefasst werden in der Validierungsphase pro Quartal durchschnittlich 2% der Daten ausgeschlossen. 1% kann dank der Imputation zurückgewonnen werden. In der Plausibilisierungsphase bzw. bei der Verarbeitung der Extremwerte werden rund 5% der Transaktionen entfernt. Insgesamt werden rund 6% der Daten aus der Berechnung des Index ausgeschlossen.

5.2 Aggregationsschritte

Sobald die Daten validiert, imputiert und plausibilisiert sind (vgl. Kapitel 5.1), werden sie nach Objekttyp und Gemeindetyp in zehn Schichten eingeteilt (vgl. Kapitel 4.2). Für jede Immobilientransaktion wird ein fiktiver Wert zu konstanten impliziten Preisen (vgl. Kapitel 4) berechnet, sodass jeder Beobachtung zwei Preise zugeordnet werden können: der effektive Transaktionspreis (Bruttopreis) und ein geschätzter Wert.

Im ersten Aggregationsschritt werden pro Zelle zwei Durchschnittspreise berechnet. Dazu wird das geometrische Mittel verwendet, das auch für die anderen Preisindizes verwendet wird. Es wird in der Preisstatistik sehr geschätzt, da es interessante mathematische Eigenschaften aufweist, beispielsweise die Transitivität[22], die etwa im Hinblick auf eine Verkettung sehr wichtig ist. Diese Durchschnittswerte werden innerhalb jeder der zehn Zellen berechnet.

$$\bar{p}_j^t = \left[\prod\nolimits_{i \in j} p_i^t\right]^{\frac{1}{n_j}} \quad (1)$$

$$\bar{p}_{est,j}^t = \left[\prod\nolimits_{i \in j} \hat{p}_i^t\right]^{\frac{1}{n_j}} \quad (2)$$

wobei:

$\bar{p}_j^t$ = geometrisches Mittel der Transaktionspreise für die Immobilien innerhalb Zelle j im Quartal t
$\bar{p}_{est,j}^t$ = geometrisches Mittel der geschätzten Preise für die Immobilien innerhalb Zelle j im Quartal t
j = Zelle (Objekttyp X Gemeindetyp)
t = Quartal t
n_j = Zahl der Transaktionen innerhalb Zelle j
p_i^t = Transaktionspreis für Immobilie i im Quartal t
$\hat{p}_i^t$ = Schätzpreis für Immobilie i im Quartal t

Im zweiten Aggregationsschritt werden Elementarindizes anhand der durchschnittlichen Transaktionspreise (Bruttopreisindex) und der durchschnittlichen geschätzten Preise (Qualitätsindex) gebildet.

$$IP_j^{0,t} = \frac{\bar{p}_j^t}{\bar{p}_j^0} \times 100 \ (3)$$

$$IQ_j^{0,t} = \frac{\bar{p}_{est,j}^t}{\bar{p}_{est,j}^0} \times 100 \ (4)$$

wobei:

$IP_j^{0,t}$ = Bruttopreisindex, berechnet mit dem geometrischen Mittel der Transaktionspreise aus Quartal t in Zelle j für einen Objekt- und Gemeindetyp
$IQ_j^{0,t}$ = Qualitätsindex, berechnet mit dem geometrischen Mittel der für das Quartal t geschätzten Preise für einen Objekt- und Gemeindetyp in Zelle j
t = aktuelles Quartal
0 = Basisquartal

Innerhalb jeder Zelle wird der Bruttopreisindex anschliessend durch den Qualitätsindex geteilt, um die Unterschiede in der Qualität zu neutralisieren.

$$IPa_j^{0,t} = \frac{IP_j^{0,t}}{IQ_j^{0,t}} \times 100 \ (5)$$

wobei:

$IPa_j^{0,t}$ = qualitätsbereinigter Preisindex pro Zelle (nach Objekt- und Gemeindetyp) in der aktuellen Periode t im Vergleich zum Basisquartal 0. Diese Berechnung erfolgt für sämtliche Zellen:

[22] Das Axiom der Transitivität geht davon aus, dass der Index zwischen 0 und N unter Verwendung der einzelnen Teilperioden N-1, N-2, N-3 berechnet werden kann.

Teilindizes	Einfamilienhäuser (EFH)	Eigentumswohnungen (EGW)
SGA – Städtische Gemeinde einer grossen Agglomeration	$IPa_{(EFH,UGA)}$	$IPa_{(EGW,UGA)}$
SMA – Städtische Gemeinde einer mittelgrossen Agglomeration	$IPa_{(EFH,UAM)}$	$IPa_{(EGW,UAM)}$
SKAA – Städtische Gemeinde einer kleinen oder ausserhalb einer Agglomeration	$IPa_{(EFH,UPHA)}$	$IPa_{(EGW,UPHA)}$
INT – Intermediäre Gemeinde	$IPa_{(EFH,INT)}$	$IPa_{(EGW,INT)}$
LAN – Ländliche Gemeinde	$IPa_{(EFH,RUR)}$	$IPa_{(EGW,RUR)}$

Mit dem dritten und letzten Aggregationsschritt kann der Wohnimmobilienpreisindex nach Objekttyp und Gemeindetyp sowie insgesamt berechnet werden. Jeder Subindex wird nach dem Gewicht seiner Zelle mittels einer Laspeyres-Formel (Young-Formel: gewichtetes arithmetisches Mittel) berücksichtigt.

$$IPa_C^{0,t} = \frac{\sum_{j\in C}\left[IPa_j^{0,t} \times g_{j,B}\right]}{\sum_{j\in C}\left[g_{j,B}\right]} (6)$$

$$IPa_O^{0,t} = \frac{\sum_{j\in O}\left[IPa_j^{0,t} \times g_{j,B}\right]}{\sum_{j\in O}\left[g_{j,B}\right]} (7)$$

$$IPa^{0,t} = \frac{\sum_j \left[IPa_j^{0,t} \times g_{j,B}\right]}{\sum_j \left[g_{j,B}\right]} (8)$$

wobei:

$IPa_C^{0,t}$ = qualitätsbereinigter Preisindex für den Gemeindetyp C im Quartal t im Vergleich zum Basisquartal 0

$IPa_O^{0,t}$ = qualitätsbereinigter Preisindex für einen Objekttyp O im Quartal t im Vergleich zum Basisquartal 0

$IPa^{0,t}$ = qualitätsbereinigter Preisindex total im Quartal t im Vergleich zum Basisquartal 0

$g_{j,B}$ = Gewicht der Zelle j im Gewichtungsjahr B (Vorjahr)

Verkettung: Das Gewicht der Zellen ($g_{j,B}$) wird jedes Jahr aktualisiert, damit die Struktur des Immobilienmarkts möglichst genau der Realität entspricht. Um mit den Ergebnissen lange Zeitreihen zu erstellen, werden die Indizes miteinander verkettet. Verkettungsquartal ist das 4. Quartal (q4), das jeweils als neues Basisquartal und folglich als Verbindungsglied zwischen dem Index auf der alten Basis und dem Index auf der neuen Basis dient.

wobei:

IPa^{q4_{T-n},t_T} = qualitätsbereinigter Preisindex im Quartal t des Jahres T im Vergleich zur Basisperiode, d.h. dem 4. Quartal des Jahres T-n

IPa^{q4_{T-1},t_T} = qualitätsbereinigter Preisindex im Quartal t des Jahres T im Vergleich zur Basisperiode, d.h. dem 4. Quartal des Jahres T-1

$IPa^{q4_{T-n},q4_{T-n+1}}$ = qualitätsbereinigter Preisindex im 4. Quartal des Jahres T-n+1 im Vergleich zur Basisperiode, d.h. dem 4. Quartal des Jahres T-n

n = Zahl der Verbindungsglieder (hier: ein Verbindungsglied = 1 Jahr) neue Basisquartale)

$$IPa^{q4_{T-n},t_T} = IPa^{q4_{T-n},q4_{T-n+1}} \times IPa^{q4_{T-n+1},q4_{T-n+2}} \times \ldots \times IPa^{q4_{T-1},t_T} \times \frac{1}{100^n} (9)$$

Zusammengefasst G 5