# Rozdział 3
# Algorytm postępowania badawczego – metody i techniki badawcze

## 3.1. Wprowadzenie

Przedstawione w poprzedniej części pracy ustalenia teoretyczne w kolejnych rozdziałach zostaną poddane empirycznej weryfikacji. Chcąc jednak przybliżyć Czytelnikowi procedury empirycznej weryfikacji, celem tego rozdziału jest prezentacja metod badawczych i technik zbierania informacji. Metody i techniki badawcze posłużyły trzem głównym celom. Po pierwsze, identyfikacji poziomu rozwoju społeczno-gospodarczego i jego dynamiki w układzie gmin Wielkopolski na tle innych tego typu jednostek w kraju. Po drugie, identyfikacji czynników rozwoju społeczno-gospodarczego na poziomie lokalnym. Po trzecie, identyfikacji trzech zarysowanych wyzwań polityki rozwoju, tj. identyfikacji kapitału terytorialnego, dyfuzji procesów rozwojowych oraz znaczenia instytucji w rozwoju, które *de facto* powinny stanowić podstawę polityki zorientowanej terytorialnie (ryc. 3.1). Analiza empiryczna prowa-


Ryc. 3.1. Ogólny algorytm postępowania badawczego

Źródło: opracowanie (dane GUS).

dzona na poziomie lokalnym pozwoli określić pozycję gmin województwa wielkopolskiego na skali poziomu i dynamiki rozwoju, ale także poznać szczegółowe uwarunkowania polityki zorientowanej terytorialnie w wybranych gminach reprezentujących wyróżnione klasy rozwoju.

Realizacja założonych celów wymaga przyjęcia określonych reguł postępowania badawczego. Ze względu na to wyróżniono trzy grupy metod i technik badawczych odpowiadające wyróżnionym celom.

Przyjęte postępowanie badawcze w pierwszej części dotyczącej analizy poziomu rozwoju społeczno-gospodarczego w ogólnej postaci ma następujący algorytm: (1) dobór zmiennych diagnostycznych, (2) wyznaczenie wskaźnika syntetycznego rozwoju społeczno-gospodarczego (poziomu rozwoju i jego dynamiki), (3) klasyfikacja gmin na podstawie wartości wskaźnika syntetycznego. Generalnie zadaniem tego algorytmu jest uzyskanie możliwie wysokiego stopnia poprawności i efektywności klasyfikacji jednostek przestrzennych na podstawie wskaźników opisujących wybrane aspekty rozwoju społeczno-gospodarczego (pochodzących z oficjalnej statystyki publicznej). Jest to o tyle ważne, że od wyników klasyfikacji w dużej mierze zależą kolejne czynności badawcze oraz ich jakość i skuteczność, a także w wielu przypadkach decyzje podejmowane przez przedstawicieli świata praktyki, w tym m.in. decyzje dotyczące koncentracji tematycznej i przestrzennej polityki rozwoju prowadzonej i realizowanej zarówno na poziomie UE, państw narodowych, regionów, jak i jednostek lokalnych [por. Nijkamp 1986; Duranton i in. 2015]. Dodatkowo część tę wzbogacono o analizę poziomu nierówności wewnątrzregionalnych w Polsce w zakresie poziomu rozwoju społeczno-gospodarczego. W części drugiej, dotyczącej identyfikacji czynników, posłużono się modelowaniem regresyjnym w ujęciu przestrzennym. W tym przypadku ogólny algorytm postępowania bazuje głównie na szacowaniu parametrów strukturalnych równań regresji, oceny stopnia dopasowania modeli oraz analizie reszt z regresji. W części trzeciej, dotyczącej identyfikacji specyficznych uwarunkowań lokalnych polityki zorientowanej terytorialnie przez identyfikację kapitału terytorialnego, dyfuzji procesów rozwojowych oraz znaczenia instytucji w rozwoju wybranych jednostek lokalnych (gmin), algorytm postępowania badawczego prezentuje się nieco odmiennie, ze względu na fakt wykorzystania wyników bezpośrednich badań ilościowo-jakościowych. Algorytm ten przedstawia się następująco: (1) wybór studiów przypadku spośród gmin województwa wielkopolskiego, (2) wykonanie wielopłaszczyznowych i pogłębionych badań terenowych o charakterze ilościowo-jakościowym w wybranych gminach województwa wielkopolskiego, (3) identyfikacja nowych wyzwań polityki regionalnej i poziomu kształtowania odporności na kryzys w badanych gminach, (4) sformułowanie rekomendacji dla zintegrowanej terytorialnie polityki rozwoju w zależności od poziomu lokalnego kapitału terytorialnego.

Realizacja celów pierwszego i drugiego, a więc analiza poziomu rozwoju społeczno-gospodarczego gmin w Polsce oraz identyfikacja czynników rozwoju, zostanie przeprowadzona w rozdziale czwartym. Natomiast realizacja celu trzeciego, zmierzająca do identyfikacji trzech wyzwań polityki rozwoju zorientowanej terytorialnie, tj. kapitału terytorialnego, dyfuzji procesów rozwojowych oraz znaczenia instytucji

w rozwoju, ze względu na jego zakres i znaczenie z punktu widzenia niniejszej pracy, zostanie przeprowadzona odpowiednio w rozdziałach piątym, szóstym, siódmym i syntetycznie w ósmym.

## 3.2. Algorytm kalsyfikacji gmin na skali poziomu rozwoju społeczno-gospodarczego

Określenie pozycji gmin w Polsce na skali poziomu rozwoju społeczno-gospodarczego, prowadzące do przyporządkowania ich do klas poziomu rozwoju oraz dynamiki rozwoju jest punktem wyjścia empirycznego postępowania badawczego. Procedura klasyfikacji gmin ma postać trójetapowego algorytmu.

W pierwszym etapie (1) postępowania badawczego dokonano wyboru i redukcji wskaźników cząstkowych opisujących wybrane aspekty rozwoju społeczno-gospodarczego. Aspekty te odpowiadają wyróżnionym w poprzednim rozdziale pięciu czynnikom rozwoju (kapitał ludzki, kapitał społeczny, kapitał materialny, kapitał finansowy i innowacje). W tym celu wykorzystano dane statystyczne publikowane w Banku Danych Lokalnych GUS dla lat 2004-2018. Rok 2004 wybrano jako moment największego (i historycznego z punktu widzenia państw zza tzw. żelaznej kurtyny) rozszerzenia UE. Taki dobór zakresu czasowego analizy pozwolił ocenić poziom rozwoju lokalnego w Polsce w momencie rozszerzenia UE. Głównym kryterium doboru wskaźników do analizy była dostępność danych w całym badanym okresie dla wszystkich jednostek przestrzennych. Pełna seria danych dla 2478 gmin w Polsce (stan na 2018 r.) była dostępna dla 38 wskaźników przyporządkowanych do pięciu aspektów rozwoju społeczno-gospodarczego odpowiadających czynnikom rozwoju. W wyniku analizy korelacji za pomocą współczynnika korelacji liniowej Pearsona (uwzględniającej interpretację merytoryczną związku liniowego między wskaźnikami a nie tylko czysty związek statystyczny) w obrębie każdego z czynników, zredukowano liczbę wskaźników do 22. Ostateczną listę wskaźników traktowanych jako zmienne diagnostyczne w kolejnych krokach postępowania badawczego zamieszczono w tab. 3.1.

W drugim etapie (2), na podstawie wartości zmiennych diagnostycznych dla każdej z badanych jednostek przestrzennych wyznaczono syntetyczny wskaźnik poziomu rozwoju społeczno-gospodarczego, syntetyczny wskaźnik dynamiki rozwoju oraz dodatkowo dla każdego z badanych pięciu aspektów syntetyczny wskaźnik poziomu ich rozwoju (kapitał ludzki, kapitał społeczny, kapitał materialny, kapitał finansowy i innowacje). Procedura wyznaczania wskaźnika syntetycznego była poprzedzona testowaniem normalności rozkładu zmiennych diagnostycznych. W większości przypadków przyjęte do analizy wskaźniki nie miały rozkładu normalnego (co potwierdziły testy normalności rozkładu zmiennej: Kołmogorowa-Smirnowa, Lillieforsa, Shapiro-Wilka i D'Agostino-Pearsona). Tym samym w celu normalizacji wskaźników nie posłużono się klasyczną metodą standaryzacji, w której wykorzystuje się średnią arytmetyczną i odchylenie standardowe, lecz metodą unitaryzacji zerowanej (normalizacją

*min-max normalization, rescaling*). Zmienne diagnostyczne będące stymulantami normalizowano za pomocą formuły (1), a zmienne będące destymulantami za pomocą formuły (2):

$$(1)\ z_{ij} = \frac{x_{ij} - \min\limits_{i} x_{ij}}{\max\limits_{i} x_{ij} - \min\limits_{i} x_{ij}};\quad (2)\ z_{ij} = \frac{\max\limits_{i} x_{ij} - x_{ij}}{\max\limits_{i} x_{ij} - \min\limits_{i} x_{ij}}$$

Tak znormalizowane wskaźniki nie tylko pozbawiają mian i sprowadzają do podobnego rzędu wielkości zmienne diagnostyczne, lecz także cechują się równością rozpiętości [0,1], możliwością normowania cech przyjmujących wartości ujemne, dodatnie i zero oraz przyjmowaniem nieujemnych wartości unormowanych [Kukuła 2000]. Następnie na bazie zunitaryzowanych wskaźników skonstruowano syntetyczny wskaźnik poziomu rozwoju społeczno-gospodarczego metodą wzorca rozwoju. W tym celu wykorzystano miarę niepodobieństwa Braya-Curtisa [Bray, Curtis 1957], którą transformowano na miarę podobieństwa do wzorca, którym była hipotetyczna jednostka przestrzenna przyjmująca dla wszystkich uwzględnianych w analizie zmiennych wartość maksymalną (tj. 1):

$$d_{kj}^{BC} = 1 - \frac{\sum_{j=1}^{m} | z_{ij} - z_{kj} |}{\sum_{j=1}^{m} (z_{ij} + z_{kj})}$$

gdzie:

$z_{ij}$ – znormalizowana wartość wskaźnika $j$ dla gminy $i$ ($i$=1, 2, …, 2478),

$k$ – gmina „wzorzec”,

$j$ = 1, 2,…, $m$ – numer wskaźnika ($m$ = 22).

Wyprowadzone w ten sposób wskaźniki syntetyczne przyjmowały wartości w przedziale [0,1], a im wyższa wartość wskaźnika syntetycznego tym wyższy poziom rozwoju społeczno-gospodarczego (oraz wyższy poziom dynamiki rozwoju i wyższy poziom rozwoju czynnika rozwoju). Tym samym dla każdej z 2478 gmin dla każdego roku objętego analizą skonstruowano odpowiednio 15 wskaźników poziomu rozwoju społeczno-gospodarczego, po 15 wskaźników poziomu rozwoju dla każdego z pięciu wyróżnionych aspektów oraz 14 wskaźników dynamiki rozwoju. W tym miejscu warto podkreślić, że ze względu na konieczność wyznaczenia względnej dynamiki badanych zjawisk (gdzie 2004=100), a część wskaźników przyjmowała wartości ujemne lub zero, dynamika ta nie była liczona na wartościach oryginalnych wskaźników, ale na ich wartościach znormalizowanych (zunitaryzowanych) względem całej serii czasowej i przesuniętych o stały wektor jednego odchylenia standardowego. W ten sposób każda z cech była początkowo unormowana w przedziale [0,1], a następnie przesunięta o wartość jednego odchylenia standardowego, co sprawiło, że pozostały jedynie liczby dodatnie, i tym samym można było na ich podstawie wyznaczyć względną dynamikę zjawiska.

W trzecim etapie (3) dokonano klasyfikacji badanych jednostek przestrzennych. Przy czym przez klasyfikację należy rozumieć z jednej strony czynność wyodrębniania

Tabela 3.1. Zmienne diagnostyczne

| Czynnik i kod | | Zmienne | Typ |
|---|---|---|---|
| Kapitał ludzki | KL_1 | ludność w wieku nieprodukcyjnym na 100 w wieku produkcyjnym | D |
| | KL_2 | przyrost naturalny na 1000 ludności | S |
| | KL_3 | saldo migracji wewnętrznych i zagranicznych na 1000 ludności | S |
| | KL_4 | przychodnie na 10 tys. ludności | S |
| | KL_5 | bezrobotni na 100 osób w wieku produkcyjnym | D |
| Kapitał społeczny | KS_1 | pracujący na 1000 osób w wieku produkcyjnym | S |
| | KS_2 | fundacje, stowarzyszenia, organizacje na 1000 lub 10 tys. osób | S |
| | KS_3 | osoby fizyczne prowadzące działalność gospodarczą na 1000 ludności | S |
| | KS_4 | udział wyższych urzędników, kierowników oraz specjalistów w ogóle radnych (%) | S |
| | KS_5 | współczynnik skolaryzacji netto gimnazjów | S |
| | KS_6 | liczba dodatków mieszkaniowych na 1000 mieszkańców | D |
| Kapitał materialny | KM_1 | udział obszarów prawnie chronionych w powierzchni gminy (%) | S |
| | KM_2 | różnica pomiędzy odsetkiem ludności korzystającej z wodociągu i z kanalizacji | D |
| | KM_3 | przeciętna powierzchnia użytkowa mieszkania na 1 osobę ($m^2$/os.) | S |
| | KM_4 | udział mieszkań posiadających ustęp spłukiwany (%) | S |
| | KM_5 | udział mieszkań posiadających podłączenie do gazu sieciowego (%) | S |
| Kapitał finansowy | KF_1 | wydatki majątkowe inwestycyjne gmin na 1 mieszkańca (zł/os.) | S |
| | KF_2 | dochody z podatku PIT na 1 mieszkańca (zł/os.) | S |
| | KF_3 | dochody z podatku na 1 mieszkańca (zł/os.) | S |
| | KF_4 | dochody własne *per capita* (zł/os.) | S |
| | KF_5 | podmioty działalności finansowej i ubezpieczeniowej na 10 tys. ludności | S |
| Innowacje | I_1 | spółki handlowe z udziałem kapitału zagranicznego na 10 tys. ludności | S |

Objaśnienia: S – stymulanta, D – destymulanta.
Źródło: opracowanie własne.

w ramach *n*-elementowego zbioru *X* takich niepustych jego *k*-podzbiorów, że spełniony jest warunek adekwatności (zupełności), a więc suma wyodrębnionych podzbiorów (klas) jest identyczna ze zbiorem wyjściowym $X_1 \cup X_1 \cup \ ... \cup X_k \ = \ X$ oraz warunek rozłączności mówiący o tym, że poszczególne podzbiory nie zawierają elementów wspólnych $X_i \cap X_j \ = \emptyset$ dla $i \neq j = 1, \ 2, \ ..., \ k$; z drugiej strony, przez klasyfikację należy rozumieć efekt wyodrębniania *k*-podzbiorów, a więc konkretny podział badanych jednostek.

W celu klasyfikacji gmin podobnych pod względem wartości syntetycznego indeksu poziomu i dynamiki rozwoju społeczno-gospodarczego zastosowano metody eksploracji danych (*data mining*) i uczenia maszynowego (*machine learning*). W pierwszym kroku zastosowano iteracyjną niehierarchiczną metodę grupowania – analizę skupień według algorytmu *k*-średnich (*cluster analysis, k-means clustering*), a w drugim zastosowano metodę weryfikującą uzyskaną klasyfikację – metodę lasów losowych (*random forest*).

Analiza skupień według algorytmu *k*-średnich należy do grupy metod określanych mianem klasyfikacji bezwzorcowej lub klasyfikacji nienadzorowanej. Metoda ta nie wymaga normalności rozkładu zmiennych, a ze względu na fakt, że każdorazowo prowadzona była na podstawie wartości jednego wskaźnika syntetycznego spełnia warunek braku współliniowości. Uproszczając można przyjąć, że celem tej wersji analizy skupień jest utworzenie *k*-niepustych, rozłącznych i względnie jednorodnych klas (tzw. skupień), w taki sposób, że w każdej iteracji część obiektów jest przenoszona między skupieniami, tak aby maksymalizować wariancje międzygrupowe i minimalizować wariancje wewnątrzgrupowe [Hartigan 1975; Kaufman, Rousseeuw 1990]. W postępowaniu badawczym przyjęto, że $k$=3. Wynikało to przede wszystkim z dwóch przesłanek: (1) chęci wydzielenia nieparzystej liczby klas (np.: wysoki, przeciętny, niski poziom rozwoju) oraz (2) zapewnienia odpowiedniej liczności poszczególnych klas (np. dla $k$=5 jedna z klas składała się tylko z kilku gmin) co w dużej mierze podyktowane było następnym krokiem postępowania badawczego. Niemniej jednak warto podkreślić, że testowano także podziały na 4, 5 i 7 klas, jednak warianty te nie zapewniały wyraźnego przyrostu wariancji międzygrupowej oraz spadku wariancji wewnątrzgrupowej w relacji do $k$=3, a więc nie sprawiało to, że wydzielone klasy były bardziej jednorodne. Ze względu na fakt, że klasyfikacja miała charakter pseudojednocechowy, gdyż odbywała się na podstawie wartości wskaźnika syntetycznego, możliwe było uporządkowanie liniowe badanych jednostek przestrzennych i deskrypcja wydzielonych klas (skupień) jako klasy: wysokiego, przeciętnego i niskiego poziomu rozwoju (dynamiki rozwoju).

Jak już wspomniano, w drugim kroku w celu polepszenia uzyskanych klasyfikacji dokonano ich weryfikacji za pomocą metody lasów losowych. W odróżnieniu od analizy *k*-średnich, metoda lasów losowych jest przykładem klasyfikacji wzorcowej lub inaczej uczenia nadzorowanego [Larose, Larose 2015]. Najogólniej rzecz ujmując celem tego typu metod bazujących na statystyce i twierdzeniu Baysea jest znalezienie reguły klasyfikacji, czyli zbudowanie formalnego modelu zwanego klasyfikatorem, na podstawie obserwacji ze zbioru uczącego, treningowego (zbioru zawierającego sklasyfikowane obserwacje, czyli takie, dla których wartość zmiennej zależnej, a więc klasa, jest znana), następnie jego testowanie na pozostałych obserwacjach skonstruowanym klasyfikatorem i przydzielanie do odpowiednich klas wskazanych przez „wyuczony" klasyfikator (etap weryfikacji dokładności, jakości klasyfikatora na podstawie zbioru testowego danych). Decyzja o przydzieleniu obserwacji do danej klasy odbywa się na podstawie rozkładu zmiennych w klasach i wielkości prawdopodobieństw *a priori*. W uproszczeniu można przyjąć, że zmienną objaśnianą w tych procedurach jest (w analizowanych przypadkach) klasa poziomu (lub dynamiki) rozwoju społeczno-

-gospodarczego (zmienna nominalna), a zmiennymi objaśniającymi określone wskaźniki cząstkowe (zmienne ilorazowe). W przypadku badanego zbioru gmin tymi zmiennymi były 22 oryginalne wskaźniki (zmienne diagnostyczne). Zatem procedura weryfikacji klasyfikacji uzyskanej w wyniku klasyfikacji pseudojednocechowej za pomocą analizy *k*-średnich weryfikowana jest przez różne kombinacje zmiennych diagnostycznych, tworzących wskaźniki syntetyczne poziomu rozwoju (dynamiki rozwoju).

W celu przybliżenia Czytelnikowi metody lasów losowych poniżej przedstawiamy jej krótką charakterystykę i ogólną ideę. Metoda ta została stworzona przez L. Breimana [2001] na przełomie XX i XXI w. i nie doczekała się jeszcze wielu zastosowań na gruncie nauk społecznych (w tym w geografii społeczno-ekonomicznej i gospodarce przestrzennej). Dlatego też w nieco szerszym ujęciu prezentujemy tę metodę. Jednak należy podkreślić, że celem autorów pracy nie jest podanie dokładnych wzorów i algorytmów obliczeniowych, stąd też Czytelników zainteresowanych matematycznymi podstawami metody odsyłamy do szczegółowych opracowań, na podstawie których powstał niniejszy opis.

W celu lepszego zrozumienia idei metody lasów losowych krótko należy scharakteryzować ideę metody drzew decyzyjnych (*decision trees*), nazywanych także drzewami klasyfikacyjnymi (*classification trees*), które stanowią podstawę omawianej metody. Do najpopularniejszych metod opartych na drzewach decyzyjnych należą m. in. algorytm C4.5, CHAID (*CHi-square Automatic Interaction Detection*), CART (*Classification And Regression Tree*), MARS (*Multivariate Adaptive Regression Spline*). Drzewo decyzyjne (które jest klasyfikatorem) jest grafem spójnym i acyklicznym. Składa się z korzenia (wierzchołka) oraz wychodzących z niego gałęzi (krawędzi) prowadzących do węzłów. W korzeniu znajdują się wybrane do próby uczącej obiekty, które przesuwane są w dół drzewa przez kolejne gałęzie do kolejnych węzłów. W każdym z węzłów podejmowana jest decyzja o wyborze gałęzi, w kierunku której przesuwane są obiekty (powstaje reguła decyzyjna). W ten sposób próba zostaje podzielona na podgrupy. Pod każdym węzłem wyszczególnione jest kryterium podziału dokonywanego w danym węźle, które jest jednakowe dla wszystkich elementów próby. Kryterium tym najczęściej jest test binarny, tzn. test o dwuelementowym zbiorze wyników. Ostatecznym etapem budowy są liście drzewa, którym przypisane zostają klasy. Ważnym zagadnieniem jest kwestia wyboru odpowiedniej reguły decyzyjnej, a więc miary różnorodności (podziału) w węźle, która będzie stanowiła podstawę klasyfikacji oraz algorytmu maksymalizacji tej różnicy, a więc algorytmu wyznaczania wartości uznanych za graniczne (stanowiących podstawę podziału). Tym samym cała procedura polega na wyborze próby uczącej „przesuwającej się" w dół do określonych węzłów, w których do wyboru reguł decyzyjnych w węźle stosuje się kryteria w postaci: ułamka błędnych klasyfikacji, indeksu Giniego lub entropii, które prowadzą do wyboru najlepszego wskaźnika i jego wartości umożliwiających podział na najbardziej jednorodne klasy (podpróby). Procedura uczenia drzewa zatrzymywana jest w momencie spełnienia kryterium zatrzymania (np. określona głębokość – liczba węzłów, w liściu pozostaną jedynie próbki pochodzące z tej samej klasy) lub występowania tylko jednej próbki w liściu. W trakcie wskazywania przez drzewo etykiety klasy dla próbki, jest ona poddawana kolejnym testom w węzłach drzewa, począwszy od korzenia. Na koniec

propagacji próbki przez drzewo, próbce jest przypisywana klasa, jaką posiada liść. Metoda drzew decyzyjnych jest dość prosta i szybka w działaniu, zapewniająca przejrzystość klasyfikacji. Niestety drzewa decyzyjne mają także kilka wad, wśród których najważniejszymi są niestabilność i niska skuteczność. Oznacza to, że nawet przy małej zmianie zbioru uczącego uzyskane reguły decyzyjne w węzłach mogą się zmienić, a przy złożonych zbiorach pojedyncze drzewa decyzyjne mogą osiągać niską skuteczność klasyfikacji.

Lasy losowe stanowią próbę eliminacji wad klasycznych binarnych drzew klasyfikacyjnych [Breiman i in. 1998]. Metoda ta jest hybrydą *baggingu* (*Bootrap AGGregatING*) i metody losowych podprzestrzeni (*Random Subspace Method*). W dużym uproszczeniu można przyjąć, że *bagging* (którego twórcą jest także L. Breiman [1996]) należy do procedur agregujących rodzinę klasyfikatorów w jeden zbiorczy klasyfikator, który wynik klasyfikacji opiera na głosowaniu większościowym. Rodzina *L* klasyfikatorów $\{\Psi_1,\ \Psi_2,\ \ldots\ \Psi_L\}$ jest generowana na podstawie *L* ciągów uczących, utworzonych przez *n*-krotne losowanie ze zwracaniem (jest to tzw. próba *bootstrapowa*) elementów ciągu uczącego o długości *N*. Wyuczone na ich podstawie klasyfikatory $\Psi_1,\ \Psi_2,\ \ldots\ \Psi_L$ podejmują wspólną decyzję o klasyfikacji obiektu do jednej z klas, gdy dana klasa uzyska najwięcej wskazań. Z kolei metoda losowych podprzestrzeni [Ho 1998; Skurichina, Duin 2002] jest podobna do *baggingu*, jednak tutaj zmienne niezależne („atrybuty", „predyktory") dobierane są losowo na zasadzie próby losowej możliwie najbardziej zróżnicowane dla każdego z obiektów w zbiorze uczącym. Chodzi więc o dodanie dodatkowego losowego doboru zmiennych niezależnych w każdej pętli próby *bootstrapowej*. Zatem najogólniej rzecz ujmując metoda lasów losowych polega na łączeniu wielu drzew klasyfikacyjnych (las) bez przycinania, na wielu losowo dobranych próbach (losowo dobierane obiekty i zmienne) z jednoczesnym podziałem zbioru na zbiór uczący i zbiór testowy, w której ostateczna klasyfikacja powstaje w wyniku „głosowania" (wybór większościowy) zespołu drzew. Podczas tworzenia poszczególnych drzew decyzyjnych zbiór wejściowy dzielony jest na dwa podzbiory: uczący (treningowy) oraz testowy (*out-of-bag* – OOB). Zbiór OOB służy do estymacji błędów klasyfikacji oraz istotności poszczególnych zmiennych. Błąd predykcji OOB wskazuje, ile elementów ze zbioru testowego nie zostało przyporządkowanych poprawnie do ich właściwych klas. Jest on różnicą między wszystkimi elementami znajdującymi się w macierzy trafności a elementami znajdującymi się poza przekątną macierzy. Zatem punktem wyjścia w algorytmie lasów losowych jest losowanie ze zwracaniem z *n*-elementowej próby uczącej *n* wektorów obserwacji (próba *bootstrapowa*). Na podstawie takiej pseudopróby tworzone jest drzewo. W każdym węźle podział odbywa się przez wylosowanie bez zwracania *m* spośród *p* wskaźników (cech), następnie w kolejnym węźle *k* spośród *m* wskaźników itd. ( $p \gg m \gg k$) (przyjmuje się, że $m \cong \sqrt{p}$ daje dobre wyniki). Budowanie drzewa bez przycinania trwa do momentu uzyskania w liściach elementów z tylko jednej klasy. Dany wektor obserwacji jest klasyfikowany przez wszystkie drzewa (głosowanie), ostatecznie zaklasyfikowany do klasy, w której wystąpił najczęściej. W przypadku elementów niewylosowanych z oryginalnej podpróby, każdy taki element zostaje poddany klasyfikacji przez drzewa, w których budowie nie brał udziału. Taki element zostaje następnie przyporządkowany klasie, która

osiągana była najczęściej (w ten sposób zaklasyfikowane zostały wszystkie elementy z oryginalnej próby). W algorytmie lasów losowych zaimplementowanym w pakiecie Statistica (wykorzystanym w niniejszej pracy) jako kryterium podziału w węźle wykorzystuje się indeks Giniego:

$$T(Gini) = 1 - \sum_{j=1}^{n} p_j^2$$

gdzie:

$T$ – zbiór zawierający $n$ klas

$p_j$ – względna częstotliwość występowania danej klasy w zbiorze $T$.

Wzrost wartości współczynnika oznacza wzrost nierówności rozkładu, tym samym wybierana jest klasa, która charakteryzuje się najniższą wartością indeksu Giniego.

Połączenie *baggingu* i metody losowych podprzestrzeni zapewnia minimalizację błędu modelu z jednoczesnym utrzymaniem relatywnie małego jego obciążenia (kwadratu różnicy między wartością oczekiwaną przewidywań modeli dla różnych prób a wartością obserwowaną) – takiego jak dla pojedynczego drzewa oraz relatywnie niskiej wariancji modelu (przez tworzenie drzew w najmniejszym stopniu skorelowanych ze sobą dzięki uczeniu drzew klasyfikacyjnych na próbach losowanych ze zwracaniem oraz przez losowanie pewnej liczby zmiennych objaśniających spośród wszystkich zmiennych przed każdym podziałem w drzewie i tylko na tych zmiennych budowaniu klasyfikacji). Zaletą metody lasów losowych jest (1) odporność na współliniowość zmiennych, wartości odstające i dużą liczbę zmiennych objaśniających, (2) możliwość odtworzenia złożonych zależności i wykrycia interakcji między zmiennymi, (3) możliwość określenia wskaźników determinujących klasyfikację oraz (4) odporność na „przeuczenie" klasyfikatora, czyli nadmiernego dopasowania się drzew do próby uczącej [Breiman 2001; Hastie i in. 2013].

Ocenę jakości klasyfikacji najczęściej prowadzi się na podstawie tzw. macierzy pomyłek i krzywych ROC [Fawcett 2006]. W przypadku macierzy pomyłek wykorzystuje się m.in. następujące miary: *sensitivity* (TPR – *true positive rate*), *specificity* (SPC) (TNR – *true negative rate*), *precision* (PPV – *positive predictive value*), NPV (*negative predictive value*), FPR (*false positive rate*) oraz *Matthews correlation coefficient* (MCC), *Markedness* (MK), *F1 score i accuracy* (ACC). Natomiast krzywa ROC (*Receiver Operating Characteristic*) jest graficzną reprezentacją efektywności modelu predykcyjnego i przedstawia zmienność TPR (miary pokrycia – wykrycia klasy faktycznie pozytywnej) w zależności od FPR (poziomu błędu popełnianego na klasie faktycznie negatywnej).

W wyniku zarysowanej powyżej procedury klasyfikacji wydzielono klasy gmin na skali poziomu rozwoju społeczno-gospodarczego (i dynamiki rozwoju). Dla każdego roku obserwacji, każda z badanych jednostek przestrzennych została przypisana do jednej z trzech klas poziomu rozwoju lub dynamiki rozwoju: wysoki, przeciętny, niski.

Ostatnim krokiem procedury klasyfikacyjnej było utworzenie klasyfikacji syntetycznej, a więc obejmującej klasyfikacje cząstkowe z całego badanego okresu. Klasyfikacja syntetyczna oparta jest na liczebnościach przynależności do jednej

z wyróżnionych klas rozwoju w całym badanym okresie. Początkowo przyjęto, że kryterium przypisania do danej klasy rozwoju (w ujęciu syntetycznym) będzie stanowiła obecność jednostki przestrzennej przez minimum 10 lat w danej klasie rozwoju (ok. 70% badanego okresu). Jednak kryterium to wymagało weryfikacji ze względu na fakt, że część gmin nie charakteryzowała się tak dużą stałością przynależności do jednej z wyróżnionych trzech klas rozwoju. W związku z tym w klasyfikacji syntetycznej ostatecznie wydzielono pięć klas. Oprócz trzech stałych klas – wysokiego, przeciętnego i niskiego rozwoju wydzielono klasę przeciętno-wysokiego i przeciętno-niskiego poziomu rozwoju. W klasach tych znalazły się te gminy, które w różnych latach należały do różnych klas rozwoju i najczęściej balansowały na granicy odpowiednio wysokiego i przeciętnego oraz przeciętnego i niskiego rozwoju.

Uzupełnieniem procedury klasyfikacji jest analiza stopnia nierówności pod względem poziomu rozwoju społeczno-gospodarczego gmin w poszczególnych województwach, a także występujących w tym zakresie zmian i tendencji. W celu analizy nierówności wykorzystano trzy klasyczne indeksy: (1) Williamsona [Williamson 1965], określany często mianem ważonego ludnością współczynnika zmienności, (2) Giniego [Dixon i in. 1987] dla szeregu uporządkowanego rosnąco oraz (3) Theila *T* [Theil 1996], w postaci:

$$(1)\ CV_w = \frac{1}{\bar{y}} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \cdot \frac{P_i}{\sum P_i}}$$

$$(2)\ G_w = \frac{1}{n^2\ \bar{y}}\ \cdot \sum_{i=1}^{n} (2i - n - 1) y_i$$

$$(3)\ T = \frac{T_T + T_L}{2}$$

$$gdzie:$$

$$T_T = GE(1) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\bar{y}}\ \ln\left(\frac{y_i}{\bar{y}}\right)$$

$$T_L = GE(0) = \frac{1}{n} \sum_{i=1}^{n}\ \ln\left(\frac{\bar{y}}{y_i}\right)$$

gdzie:

$y_i$ – wartość analizowanego wskaźnika dla jednostki *i*,
$\bar{y}$ – wartość średnia analizowanego wskaźnika,
$P_i$ – wielkość populacji jednostki *i*
*n* – liczba analizowanych jednostek przestrzennych.

Indeksy Williamsona (1) i Giniego (2) są *de facto* wskaźnikami konwergencji typu sigma, a więc wskaźnikami, które pozwalają stwierdzić czy między badanymi jednostkami przestrzennymi różnice w czasie się zmniejszają czy też zachodzi sytuacja odwrotna. Pewną „wyższością" indeksu Giniego nad indeksem Williamsona, jest jego unormowanie w przedziale [0,1]. Z kolei indeks Theila (3) bazuje na pojęciu entropii, a więc braku uporządkowania i zaliczany jest do tzw. wskaźników uogólnionej entropii. Przyjmuje wartości w przedziale od 0 do ln(n) (co raczej jest wartością teoretyczną), a najczęściej nie przekracza 1. We wszystkich przypadkach wartość 0 oznacza brak nierówności, wzrost ich wartości oznacza wzrost nierówności.

## 3.3. Algorytm identyfikacji czynników rozwoju społeczno-gospodarczego

Drugim etapem realizowanego w prezentowanej pracy postępowania badawczego jest identyfikacja czynników rozwoju społeczno-gospodarczego na poziomie lokalnym. Zgodnie z ustaleniami z rozdziału drugiego za czynniki rozwoju przyjęto pięć wyróżnionych aspektów rozwoju społeczno-gospodarczego: kapitał ludzki, kapitał społeczny, kapitał materialny, kapitał finansowy i innowacje.

Zgodnie z regułą W. Toblera [1970: 236] *everything is related to everything else, but near things are more related than distant things* (*wszystkie obiekty są ze sobą powiązane, ale siła tych powiązań maleje wraz ze wzrostem odległości między nimi*) przyjęto założenie, że w przypadku jednostek lokalnych z dużo większym prawdopodobieństwem niż w przypadku jednostek wyższego rzędu (większych powierzchniowo i ludnościowo), dochodzi do przenikania się i dyfuzji zjawisk społeczno-gospodarczych, a więc zachodzą interakcje przestrzenne. Istnienie tych interakcji, czynnika przestrzennego, odgrywa coraz większą rolę w opisie i wyjaśnianiu zjawisk społeczno-gospodarczych. Stąd też w procedurze poszukiwania czynników rozwoju posłużono się modelami regresji przestrzennej. Takie podejście wynikało z faktu analizowania dużej liczby relatywnie niewielkich jednostek przestrzennych położonych blisko siebie. Dlatego przestrzenne modelowanie ekonometryczne uwzględniające interakcje przestrzenne wydaje się być dobrym uzupełnieniem klasycznych modeli regresji, a czasami okazuje się lepszym rozwiązaniem ze względu na wyższą efektywność.

W modelowaniu przestrzennym punktem wyjścia jest estymacja liniowych modeli metodą najmniejszych kwadratów. Tym samym, procedura analityczna zmierzająca do wykazania związków między poziomem rozwoju społeczno-gospodarczego a czynnikami rozwoju pierwotnie bazowała na liniowych modelach regresji prostej [Maddala 2001] w postaci:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon i = 1, 2, \ldots, n$$

gdzie:

$Y_i$ – *i*-ta obserwacja zmiennej zależnej

$X_i$ – *i*-ta obserwacja zmiennej niezależnej

$\varepsilon$ – składnik losowy
$\beta_0$ – wyraz wolny (punkt przecięcia)
$\beta_1$ – parametr strukturalny modelu (współczynnik kierunkowy prostej)

Zmienną zależną (*Y*) była wartość wskaźnika syntetycznego poziomu rozwoju, a zmienną objaśniającą (niezależną) (*X*) były za każdym razem wartości wskaźników syntetycznych dla czynników (o czym wspomniano w punkcie 2 omawianego algorytmu). Dla każdego roku (lata analizy 2004-2018) wyznaczono po pięć równań regresji prostych dla wszystkich gmin jednocześnie. Parametry strukturalne modelu regresji każdorazowo szacowano klasyczną metodą najmniejszych kwadratów (*OLS – Ordinary Least Squares*). Procedura ta daje oszacowania, w której następuje dopasowanie zależności funkcyjnej do zbioru obserwacji w postaci warunku:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = min$$

gdzie suma kwadratów reszt (*SSE – Sum of the Squared Errors*), a więc suma kwadratów odchyleń wartości rzeczywistych ($y_i$) od wartości estymowanych z równania regresji ($y_i$) jest najmniejsza z możliwych. Przydatne w interpretacji wyników równań regresji oprócz wartości oszacowanych parametrów strukturalnych są także: wartość współczynnika determinacji $R^2$, wielkość testów istotności (test *t* dla poszczególnych zmiennych niezależnych oraz przede wszystkim test *F* dla całego modelu), wielkość błędu standardowego szacunku (*SEE – Standard Error of Estimates*) – ocena odchylenia standardowego składnika losowego (im mniejsza – bliższa 0, tym bardziej poprawny model), a także analiza standaryzowanych reszt z regresji. Należy podkreślić, że współczynnik determinacji w przypadku regresji prostej może mieć dwojakie znaczenie. Można rozumieć go, jako procent wyjaśnienia całkowitej zmienności zmiennej zależnej przez zmienność zmiennej niezależnej, a tym samym im jego wartość bliższa jest 1, tym większa siła wpływu zmiennej niezależnej na zmienną zależną. Wynika to też z faktu, że standaryzowany współczynnik regresji ($\beta$), który umożliwia porównania równań regresji w zakresie siły oddziaływania zmiennej niezależnej jest *de facto* równy wielkości wartości współczynnika korelacji liniowej Pearsona między zmienną zależną i zmienną niezależną. Zatem zachodzi relacja, że $\beta_1 = r_{XY}$. Z kolei w przypadku tego typu regresji prostej wielkość korelacji między zmienną zależną i niezależną podniesiona do potęgi drugiej daje wartość współczynnika determinacji. Dlatego też w celu wskazania zmiennej niezależnej (czynnika), która silniej oddziałuje na zmienną zależną (poziom rozwoju społeczno-gospodarczego) można posłużyć się wielkością $R^2$, gdyż uszeregowanie czynników pod względem wielkości tego parametru, pozwoli otrzymać takie samo uszeregowanie, jak pod względem wartości $\beta$, gdyż jak wykazano powyżej $\beta_1 = \sqrt{R^2}$. Z kolei analiza standaryzowanych reszt z regresji, wyznaczanych w postaci:

$$e_i^* = \frac{y_i - \hat{y}_\iota}{\sqrt{MSE}}$$

gdzie:

$$\sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_\iota)^2}{n-2}}$$

pozwala sprawdzić: czy słuszne jest założenie o normalności rozkładu składnika losowego?, czy wariancja składnika losowego jest stała (rozproszenie danych wokół linii regresji jest jednostajne)? – problem heteroskedastyczności, czy pominięto pewne zmienne, które powinny być włączone do modelu? czy reszty są skorelowane? (test Durbina-Watsona) – to zaś może wskazywać na wpływ relacji przestrzennych. W praktyce dodatkowo interpretacja reszt z regresji w ujęciu przestrzennym pozwala wskazać jednostki, w których występuje „nadmiar" (reszty dodatnie) lub „niedobór" (reszty ujemne) zmiennej empirycznej względem oszacowanych wartości w modelu. Oznacza to, że przy dodatnich resztach zmienna objaśniająca posiada relatywnie niskie wartości, a przy ujemnych odwrotnie. Sytuacja ta pozwala przypuszczać, że w przypadku jednostek o wysokich wartościach dodatnich lub ujemnych reszt z regresji, zmienna objaśniająca w ograniczonym stopniu pozwala objaśnić zmienność zmiennej zależnej. Jednak w związku z tym, że dla badanego układu gmin każdorazowo dla każdego z 15 badanych lat oszacowano po pięć równań regresji (dla każdego z analizowanych czynników) otrzymano 75 równań regresji i tyle samo reszt z regresji. W związku z dużym podobieństwem rozkładów przestrzennych wartości standaryzowanych reszt z regresji między kolejnymi latami, postanowiono uśrednić reszty z całego badanego okresu i analizować ich rozkład na mapach w postaci wartości średnich. Zabieg ten pozwoli na ogólny ogląd różnic między badanymi gminami w zakresie odchyleń wartości empirycznych i teoretycznych, a więc na lepszą interpretację uzyskanych wyników modelowania regresyjnego.

Ze względu na przyjęte założenie o możliwości występowania interakcji przestrzennych podjęto także badanie reszt z regresji pod kątem występowania autokorelacji przestrzennej. Autokorelację przestrzenną należy rozumieć jako korelację wartości zmiennej w danej jednostce przestrzennej z wartością analogicznej zmiennej w innej jednostce przestrzennej, która wynika z wzajemnego położenia (bliskości) tych jednostek (stymulująco wpływa ich wzajemne sąsiedztwo, a destymulująco dzieląca je odległość). W tym celu wykorzystuje się stosowną powszechnie globalną statystykę *I* Morana w postaci [Cliff, Ord 1973]:

$$I = n \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \, (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

gdzie:

$x_i$ – wartość zmiennej *x* obserwowana w jednostce *i* (*i* = 1, 2, …,*n*),

$\bar{x}$ – średnia arytmetyczna zmiennej *x* ze wszystkich *n* obserwacji,

$w_{ij}$ – element macierzy wag przestrzennych[27] standaryzowanej rzędami do jedności (każdy z jej elementów przyjmuje wartości z przedziału [0,1], a suma elementów każdego wiersza jest równa 1).

Istotność statystyczną statystyki *I* Morana, gdzie $H_0$: zmienna *x* ma rozkład losowy (brak autokorelacji przestrzennej), weryfikuje się za pomocą unormowanej statystyki $Z_I$ o rozkładzie normalnym z $E(I) = \frac{-1}{n-1} = 0$ oraz $Var(I) = 1$, w postaci:

$$Z_I = \frac{I - E(I)}{\sqrt{Var(I)}}$$

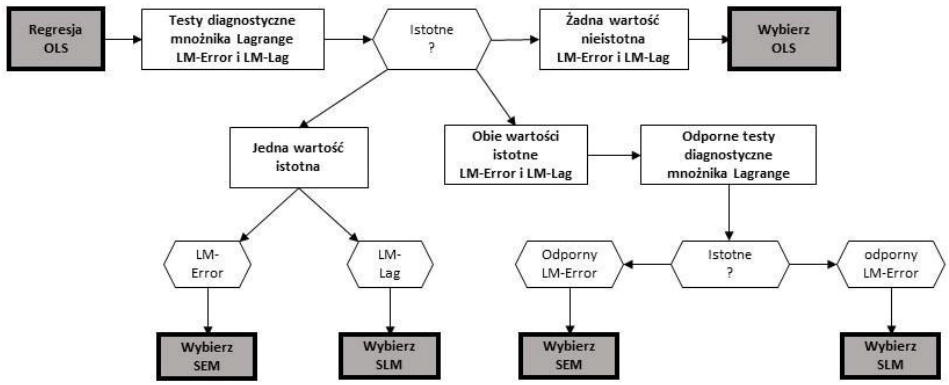
Wartości $I > \frac{-1}{n-1}$ (*de facto* dodatnie) i $Z_I > 0$ wskazują na dodatnią autokorelację przestrzenną, czyli podobieństwo sąsiadujących ze sobą jednostek w zakresie analizowanej zmiennej *x* (występują klastry wartości podobnych – wysokich lub niskich). Wartości $I < \frac{-1}{n-1}$ (*de facto* ujemne) i $Z_I < 0$ wskazują na ujemną autokorelację przestrzenną, a więc wysokie wartości zmiennej sąsiadują z niskimi (występują tzw. *hot spots*, czyli wyspy odmiennych wartości od otoczenia). Jeśli $I \approx \frac{-1}{n-1}$ (*de facto* bliskie 0) i $Z_I \approx 0$ oznacza brak autokorelacji przestrzennej badanej zmiennej, czyli wartości te rozmieszczone są losowo i niezależnie w przestrzeni [Goodchild 1986]. Uzupełnieniem globalnej statystyki Morana *I* jest jej wersja lokalna obrazująca uwarunkowania lokalne związków przestrzennych i identyfikację przestrzennych efektów aglomeracji (LISA – *Local Indicators of Spatial Associactions*). Lokalna statystyka Morana $I_i$ wskazuje, czy dana jednostka przestrzenna otoczona jest przez jednostki o podobnych lub różnych wartościach badanej zmiennej w stosunku do losowego rozmieszczenia tych wartości w przestrzeni. Lokalna statystyka Morana $I_i$ ma postać:

$$I_i = \frac{(x_i - \bar{x}) \sum_{i=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

Występowanie autokorelacji przestrzennej, zwłaszcza dodatniej, może świadczyć o nieuwzględnieniu odpowiednich zmiennych objaśniających w modelu, a tym samym może prowadzić do heterogeniczności przestrzennej. Dodatkowo świadczy to o obciążeniu i niezgodności estymatorów OLS, gdyż dane są nielosowe i występuje zła specyfikacja modelu.

Tym samym w drugim kroku należy sprawdzić za pomocą testów diagnostycznych mnożnika Lagrange (*LM – Lagrange Multiplier)* czy lepszą specyfikacją modelu ekonometrycznego będzie model opóźnienia przestrzennego (*SLM – Spatial Lag Model)* czy model błędu przestrzennego (*SEM – Spatial Error Model)* (ryc. 3.2.). W tym celu

[27] Macierz wag przestrzennych jest kwadratową macierzą sąsiedztwa o wymiarach n x n, której elementy odzwierciedlają istniejącą strukturę przestrzenną. Element przyjmuje wartość 1, jeśli i oraz j sąsiadują ze sobą, a 0 gdy nie sąsiadują. W najprostszym przypadku o sąsiedztwie decyduje wspólna granica pomiędzy badanymi jednostkami i oraz j [Anselin 1988; Kopczewska 2011]. W niniejszym badaniu przyjęto właśnie taką postać macierzy sąsiedztwa – sąsiedztwo I rzędu, najbliższy sąsiad, z którym badana jednostka posiada wspólną granicę.




Ryc. 3.2. Schemat decyzyjny przy wyborze modeli regresji przestrzennej

Źródło: opracowanie własne na podstawie [Anselin 2005: 199].

wykorzystuje się dwa zwykłe testy *LM* – $LM_{LAG}$ i $LM_{ERR}$ oraz dwa odporne – $RLM_{LAG}$ i $RLM_{ERR}$. Test $LM_{LAG}$ bada istotność przestrzennie opóźnionej zmiennej zależnej, a test $LM_{ERR}$ testuje zależność przestrzenną błędu (reszty). Odporne wersje tych testów pozwalają wykryć lokalną złą specyfikację modelu, a więc istnienie błędu przestrzennego w przypadku, gdy testowane jest istnienie opóźnienia przestrzennego i odwrotnie [Anselin i in. 1996].

Model opóźnienia przestrzennego (SLM) zawiera tzw. opóźnioną przestrzennie zmienną zależną (objaśnianą) – *spatial lag,* jako nową zmienną niezależną (objaśniającą). Zmienna ta, szacowana w postaci parametru $\rho$ jest średnią ważoną (zgodnie z przyjętą macierzą sąsiedztwa) wartości zmiennej zależnej w sąsiednich jednostkach przestrzennych. Zatem jest to model typu autoregresyjnego (*SAR – Spatial Autoregressive Models*) w postaci:

$$y_r = \rho \left( \sum_{s=1}^{n} w_{rs} y_s \right) + \sum_{i=1}^{k} \beta_i x_{ir} \; + \; \varepsilon_r$$

gdzie:

$\rho \left( \sum_{s=1}^{n} w_{rs} y_s \right)$ – *spatial lag*, określa wpływ wartości zmiennych zależnych z sąsiadujących s-tych lokalizacji na wartość tej zmiennej w danej r-tej lokalizacji.

W modelu błędu przestrzennego (SEM) uwzględnia się zależność przestrzenną składnika losowego (ogólny schemat liniowy autokorelacji przestrzennej składnika losowego) w postaci:

$$y_r = \sum_{i=1}^{k} \beta_i x_{ir} + \varepsilon_r$$

$$\varepsilon_r = \lambda \left( \sum_{s=1}^{n} w_{rs} \varepsilon_s \right) + u_r$$

gdzie:

$\varepsilon_r$ – oryginalny składnik losowy z autokorelacją przestrzenną (reszta z OLS dla *r*-tej lokalizacji), który jest funkcją przestrzennie opóźnionego błędu losowego $\sum_{s=1}^{n} w_{rs}\varepsilon_s$ (reszt z regresji z sąsiadujących *s*-tych lokalizacji) oraz „oczyszczonego" składnika losowego $u_r$ (spełniającego założenia OLS).

$\lambda$ – współczynnik siły wzajemnego skorelowania reszt z regresji OLS, pozwala wnioskować o istnieniu istotnych czynników wpływających na zmienność zmiennej zależnej, które nie zostały ujęte w modelu regresji (czynników niemierzalnych lub trudno mierzalnych, przypadkowych itp.) i określa średni wpływ sąsiednich jednostek na zmianę wielkości reszt).

Porównanie mocy eksplanacyjnej modeli OLS z modelami SLM i SEM nie powinno odbywać się na podstawie współczynnika determinacji ($R^2$), gdyż w modelach przestrzennych parametry szacuje się m.in. metodą największej wiarygodności. Stąd też ocena dopasowania modelu odbywa się na podstawie (1) logarytmicznego wskaźnika wiarygodności (*log likelihood*) oraz (2) kryteriów informacyjnych (*AIC – Akaike Information Criterion, BIC – Bayesan Information Criterion, SC – Schwartz Criterion*). W postępowaniu posłużono się wartościami logarytmu wiarygodności (najlepszym dopasowaniem odznacza się model o największej wartości) oraz AIC (najlepszy model posiada najmniejsze wartości). Dodatkowo sytuacja ta sprawia, że nie można posłużyć się wartością współczynnika determinacji jako kwadratu standaryzowanego współczynnika $b_1$, czyli $\beta_1$ (co sygnalizowano powyżej). Zatem w interpretacji siły wpływu czynnika należy posługiwać się wielkością $b_1$. W modelach przestrzennych testowanie parametrów odbywa się za pomocą testu Walda, jakości modelu za pomocą testu LR (Likelihood Ratio), testu LM oraz Morana I do testowania autokorelacji reszt. O prawidłowej specyfikacji modelu świadczą zależności: (1) dla SLM $Wals(\rho) \geq LR_{LAG} \geq LM_{LAG}$, (2) dla SEM $Wald(\lambda) \geq LR_{ERR} \geq LM_{ERR}$[28]. Tym samym wykonano po 75 estymacji modeli regresyjnych (15 lat x 5 czynników) dla OLS, SEM i SLM, co łącznie skutkowało wyznaczeniem 225 równań regresji. W analizie posłużono się wynikami modeli OLS oraz SEM, gdyż na ich zastosowanie wskazywały przeprowadzone testy LM.

## 3.4. Algorytm analizy polityki zorientowanej terytorialnie

W celu identyfikacji specyficznych uwarunkowań lokalnych polityki zorientowanej terytorialnie poprzez identyfikację kapitału terytorialnego, dyfuzji procesów rozwojowych oraz znaczenia uwarunkowań instytucjonalnych w rozwoju wybranych jednostek lokalnych (gmin), wykorzystano grupę metod i technik badań ilościowo-jakościowych.

W pierwszym etapie (1) postępowania badawczego dotyczącego badań rozwoju społeczno-gospodarczego i identyfikacji kapitału terytorialnego na poziomie lokalnym

[28] Szczegółowy opis procedury estymacji modeli przestrzennych z wykorzystaniem oprogramowania GeoDa zawiera praca L. Anselin [2005].

w województwie wielkopolskim, dokonano wyboru kilku gmin testowych, stanowiących obszar pogłębionych badań terenowych. Wyboru dokonano zgodnie z podejściem celowym – arbitralnym [Churchill 2002, Babbie 2007], biorąc pod uwagę następujące kryteria: przynależność gminy do różnych typów administracyjnych i funkcjonalnych, zróżnicowany poziom rozwoju społeczno-gospodarczego i różne położenie przestrzenne w Wielkopolsce.

W drugim etapie (2) w wybranych gminach (studiach przypadku) przeprowadzono wielopłaszczyznowe i pogłębione badania terenowe o charakterze ilościowo-jakościowym. Celem tych badań było pozyskanie opinii mieszkańców, przedsiębiorców, władz lokalnych i liderów lokalnych na temat stanu i zmian czynników rozwoju oraz wyzwań polityki regionalnej, a także możliwości i barier dla ich oddziaływania w procesie kształtowania kapitału terytorialnego badanej jednostki. W procedurze terenowych badań ilościowych i jakościowych zastosowano następujące techniki badawcze [Konecki 2000; Frankfort-Nachmias, Nachmias 2001; Babbie 2007]: wywiad kwestionariuszowy prowadzony drogą telefoniczną, wywiad pogłębiony, badania fokusowe. Wywiad kwestionariuszowy prowadzony drogą telefoniczną – CATI (*computer-assisted telephone interviewing*), jeden ze sposobów zbierania informacji, zwłaszcza o charakterze ilościowym. Wywiad telefoniczny jest prowadzony w podobny sposób co wywiad kwestionariuszowy, ale za pośrednictwem telefonu. Badanie przeprowadzane jest za pomocą specjalnego programu, który asystuje osobie dzwoniącej do respondenta przez cały czas trwania rozmowy. Program taki działa w sposób automatyczny, tzn. teleankieter na bieżąco, podczas rozmowy nanosi odpowiedzi respondentów na gotowy scenariusz wyświetlający się na ekranie. Metoda pozwala na zapisywanie wyników badania w jednym miejscu, w formie elektronicznej, bez obawy o utratę danych lub dostanie się wyników w niepowołane ręce, a koszty tego typu badania są niższe w stosunku do tradycyjnych badań ankietowych. Z kolei wywiad pogłębiony – IDI (*in-depth interview*), stanowi metodę pozyskiwania danych, głównie jakościowych, na podstawie swobodnej rozmowy z pojedynczym respondentem, np. przedstawicielem władzy lokalnej. Rozmowa dotyczy określonego tematu i prowadzona jest często z wykorzystaniem specjalnie przygotowanego scenariusza oraz nagrywana za pomocą dyktafonu. Metoda ta pozwala badać opinie zarówno z perspektywy respondenta, jak i realizatorów danego projektu, daje możliwość poruszenia skomplikowanych i szczegółowych kwestii oraz doprecyzowania lub wyjaśnienia odpowiedzi respondenta. Natomiast badania focusowe – FDI (*focus group interview*) są metodą zbierania danych, głównie jakościowych, polegającą na rozmowie specjalnie dobranej grupy osób, np. liderów lokalnych z doświadczonym moderatorem, który ukierunkowuje przebieg dyskusji na zagadnienia, które mają zostać omówione podczas spotkania. Spotkanie prowadzone jest na podstawie scenariusza, a rozmowa nagrywana jest za pomocą dyktafonu i/lub wideokamery, a następnie podlega transkrypcji. Metoda umożliwia zebranie razem różnych typów osób zaangażowanych w badanie, co daje możliwość spojrzenia na tę samą kwestię z perspektywy różnych uczestników. Wywiad kwestionariuszowy został skierowany do mieszkańców i przedsiębiorców danej jednostki terytorialnej, a wywiad pogłębiony przeprowadzono z przedstawicielem władzy lokalnej (prezydentem, burmistrzem lub wójtem). Badania focusowe wykonano z udziałem

liderów lokalnych reprezentowanych głównie przez przedstawicieli lokalnych organizacji społecznych i instytucji oraz przedstawicieli władz lokalnych niższego poziomu (sołtys) i lokalnych radnych. Zasadniczo ilościowo-jakościowe badania terenowe prowadzone były z uwzględnieniem układu pięciu obszarów badawczych tożsamych z przyjętą w niniejszej pracy klasyfikacją czynników rozwoju. Kwestionariusz ankiety przeprowadzonej wśród mieszkańców i przedsiębiorców składał się z metryki respondenta oraz zasadniczej części nawiązującej do obszarów badawczych obejmującej pytania otwarte i zamknięte w postaci kafeterii odpowiedzi (od 3 do 5 możliwości odpowiedzi, wybór tylko jednej). W kwestionariuszu zastosowano metodę pięciostopniowej skali Likerta i przyjęto założenie, że wartość 1 oznacza najsłabsze natężenie danej cechy, a wartość 5 najmocniejsze [Likert 1932]. Scenariusz wywiadu pogłębionego i badań fokusowych, podobnie jak kwestionariusz ankiety, obejmował pięć obszarów badawczych odnoszących się do analizowanych czynników rozwoju, opisanych przez pytania cząstkowe oraz pytanie dotyczące scenariusza rozwoju gminy z uwzględnieniem znaczenia poszczególnych czynników rozwoju. Dodatkowo w scenariuszu badań fokusowych zawarto metrykę respondentów[29]. W badaniu ankietowym mieszkańców podstawą wyboru wielkości próby była wielkość populacji, tj. liczba ludności oraz struktura wieku i płci w badanych jednostkach terytorialnych. Kwestionariusz ankiety mieszkańców przeprowadzono łącznie na próbie 1650 respondentów, w tym osoby w wieku produkcyjnym stanowiły 64% ogółu ankietowanych. Nieznaczną przewagę respondentów stanowili mężczyźni. W poszczególnych jednostkach terytorialnych liczba uzyskanych ankiet wynosiła od 100 (gminy wiejskie Bralin i Szczytniki) do 300 (miasto Leszno) (tab. 3.2).

Podstawą wyboru wielkości próby w badaniu ankietowym przedsiębiorców była wielkość populacji, tj. liczba przedsiębiorstw oraz ich struktura wielkościowa w badanych jednostkach terytorialnych. Kwestionariusz ankiety przedsiębiorców przeprowadzono łącznie na próbie 1350 respondentów, w tym pracujący w mikroprzedsiębiorstwach (do 9 pracujących) stanowili 93% ogółu ankietowanych. W poszczególnych jednostkach terytorialnych liczba uzyskanych ankiet wynosiła od 75 (gminy wiejskie Bralin i Szczytniki) do 400 (miasto Leszno) (tab. 3.2). Wywiady pogłębione w badanych gminach przeprowadzono bezpośrednio z urzędującymi przedstawicielami władzy lokalnej: wójtem (gminy wiejskie Bralin i Rokietnica), burmistrzem (gminy miejskie: Czarnków, Turek i Złotów), zastępcą burmistrza (gminy miejsko-wiejskie Książ Wlkp. i Trzemeszno), sekretarzem gminy (gmina wiejska Szczytniki) i incydentalnie z byłym wieloletnim prezydentem (gmina miejska Leszno) ze względu na brak możliwości przeprowadzenia wywiadu z obecnie urzędującymi przedstawicielami władzy. Badania focusowe wykonano w grupie tzw. liderów lokalnych z wyłączeniem głównych przedstawicieli władz lokalnych, reprezentowanych przez m.in. przez

[29] Przeprowadzenie wywiadów kwestionariuszowych i badań fokusowych zlecono firmie zewnętrznej: Centrum Badań Stosowanych Ultex Ankieter z Poznania, wywiady pogłębione w wybranych gminach województwa wielkopolskiego wykonali samodzielnie autorzy książki. Zakres czasowy badań terenowych zawierał się w okresie od lutego do maja 2019 r. Zakres przestrzenny badań obejmował dziewięć gmin testowych, których wykaz znajduje się w rozdziale 4.3.

Tabela 3.2. Kwestionariusz ankiety mieszkańców i przedsiębiorców – struktura próby

| Gmina | Liczba ankiet | |
|---|---|---|
| | mieszkańcy | przedsiębiorcy |
| Bralin (gm. wiejska) | 100 | 75 |
| Czarnków (gm. miejska) | 150 | 100 |
| Książ Wlkp. (gm. miejsko-wiejska) | 150 | 100 |
| Leszno (gm. miejska) | 300 | 400 |
| Rokietnica (gm. wiejska) | 200 | 100 |
| Szczytniki (gm. wiejska) | 100 | 75 |
| Trzemeszno (gm. miejsko-wiejska) | 200 | 100 |
| Turek (gm. miejska) | 250 | 200 |
| Złotów (gm. miejska) | 200 | 150 |
| **Ogółem** | **1650** | **1350** |

Źródło: opracowanie własne.

przedstawicieli organizacji społecznych: gospodarczych, sportowych i kulturalnych, dyrektorów i pracowników instytucji i przedsiębiorstw, lokalnych dziennikarzy i innych aktywistów, np. sołtysów reprezentujących władze lokalne niższego poziomu i lokalnych radnych. W przeprowadzonych wywiadach zogniskowanych wzięło udział od 7 (Leszno, Trzemeszno) do 11 liderów lokalnych (Bralin). Łącznie w wywiadach uczestniczyło 76 liderów lokalnych w wieku od 26 do 79 lat.

Postępowanie badawcze w następnym (3) etapie miało na celu zidentyfikowanie nowych wyzwań polityki regionalnej oraz poziomu odporności na kryzys, wynikających ze specyfiki oddziaływania czynników rozwoju. Podstawą wykonanej analizy są ustalenia teoretyczne zawarte w rozdz. 2 dotyczące charakterystyki trzech wyzwań polityki regionalnej: kształtowania kapitału terytorialnego, dyfuzji rozwoju i uwarunkowań instytucjonalnych procesu rozwoju. Analiza kapitału terytorialnego była prowadzona przez pryzmat czynników rozwoju z uwzględnieniem ich aspektów (subczynników). W pierwszym kroku identyfikowano ogólny poziom danego czynnika wykorzystując wskaźniki syntetyczne bazujące na odpowiednich danych wtórnych (tab. 3.1.). W dalszej kolejności badano stan poszczególnych subczynników z wykorzystaniem wskaźników strukturalnych opisujących opinie mieszkańców i przedsiębiorców gmin testowych, biorących udział w wywiadzie kwestionariuszowym (CATI). Dodatkowo uzyskane wyniki były weryfikowane wypowiedziami przedstawicieli władzy lokalnej oraz lokalnych liderów społecznych, pozyskanymi za pomocą wywiadów pogłębionych i badań fokusowych. Umożliwiło to identyfikację relacji, o charakterze stymulującym lub ograniczającym procesy rozwojowe, między analizowanym czynnikiem a pozostałymi składowymi kapitału terytorialnego. Z kolei w odniesieniu do dyfuzji rozwoju i uwarunkowań instytucjonalnych procesu rozwoju prowadzi się analogiczne postępowanie badawcze. Analiza była prowadzona przez

pryzmat czynników rozwoju z uwzględnieniem ich aspektów (subczynników). Następnie w ramach każdego czynnika ustalono istotny przejaw wyzwania w zakresie tworzenia warunków do dyfuzji rozwoju oraz wzmacniania uwarunkowań instytucjonalnych rozwoju, który powiązany jest z subczynnikiem przedmiotowego czynnika. Na tej podstawie przeprowadzono analizę empiryczną, w której wykorzystano wyniki uzyskane w ramach pogłębionych wywiadów z przedstawicielami lokalnej władzy i badań fokusowych z liderami lokalnymi (IDI, FDI), uzupełnione w miarę możliwości danymi statystycznymi (GUS). Na tej podstawie dokonano oceny cząstkowej (w układzie każdego czynnika i przejawu) i całościowej (dla wszystkich czynników i przejawów) wyzwań w przedmiotowym zakresie, a także określono poziom przygotowania gmin testowych do wzmacniania dyfuzji rozwoju i wdrażania uwarunkowań instytucjonalnych rozwoju jako wyzwań polityki regionalnej. W postępowaniu badawczym na tym etapie uwzględniono prawidłowości wynikające z przynależności gminy do danego typu administracyjno-funkcjonalnego i poziomu rozwoju społeczno-gospodarczego.

Ostatnim (4) etapem przyjętego algorytmu badawczego było sformułowanie rekomendacji dla zorientowanej terytorialnie polityki rozwoju dla poszczególnych terytoriów zróżnicowanych pod względem poziomu lokalnego kapitału terytorialnego oraz pod względem stopnia oddziaływania nowych wyzwań, przed którymi staje regionalna polityka rozwoju.